I

Overview of Mechatronics

1

What is Mechatronics?

Robert H. Bishop
The University of Texas at Austin

M. K. Ramasubramanian
North Carolina State University

Mechatronics is a natural stage in the evolutionary process of modern engineering design. The development of the computer, and then the microcomputer, embedded computers, and associated information technologies and software advances, made mechatronics an imperative in the latter part of the twentieth century. Standing at the threshold of the twenty-first century, with expected advances in integrated bio-electro-mechanical systems, quantum computers, nano- and pico-systems, and other unforeseen developments, the future of mechatronics is full of potential and bright possibilities.

1.1 Basic Definitions

The definition of mechatronics has evolved since the original definition by the Yasakawa Electric Company. In trademark application documents, Yasakawa defined mechatronics in this way [1,2]:

> The word, mechatronics, is composed of "mecha" from mechanism and the "tronics" from electronics. In other words, technologies and developed products will be incorporating electronics more and more into mechanisms, intimately and organically, and making it impossible to tell where one ends and the other begins.

The definition of mechatronics continued to evolve after Yasakawa suggested the original definition. One oft quoted definition of mechatronics was presented by Harashima, Tomizuka, and Fukada in 1996 [3]. In their words, mechatronics is defined as

> the synergistic integration of mechanical engineering, with electronics and intelligent computer control in the design and manufacturing of industrial products and processes.

That same year, another definition was suggested by Auslander and Kempf [4]:

> Mechatronics is the application of complex decision making to the operation of physical systems.

Yet another definition due to Shetty and Kolk appeared in 1997 [5]:

> Mechatronics is a methodology used for the optimal design of electromechanical products.

More recently, we find the suggestion by W. Bolton [6]:

> A mechatronic system is not just a marriage of electrical and mechanical systems and is more than just a control system; it is a complete integration of all of them.

All of these definitions and statements about mechatronics are accurate and informative, yet each one in and of itself fails to capture the totality of mechatronics. Despite continuing efforts to define mechatronics, to classify mechatronic products, and to develop a standard mechatronics curriculum, a consensus opinion on an all-encompassing description of "what is mechatronics" eludes us. This lack of consensus is a healthy sign. It says that the field is alive, that it is a youthful subject. Even without an unarguably definitive description of mechatronics, engineers understand from the definitions given above and from their own personal experiences the essence of the *philosophy* of mechatronics.

For many practicing engineers on the front line of engineering design, mechatronics is nothing new. Many engineering products of the last 25 years integrated mechanical, electrical, and computer systems, yet were designed by engineers that were never formally trained in mechatronics *per se*. It appears that modern concurrent engineering design practices, now formally viewed as part of the mechatronics specialty, are natural design processes. What is evident is that the study of mechatronics provides a mechanism for scholars interested in understanding and explaining the engineering design process to define, classify, organize, and integrate many aspects of product design into a coherent package. As the historical divisions between mechanical, electrical, aerospace, chemical, civil, and computer engineering become less clearly defined, we should take comfort in the existence of mechatronics as a field of study in academia. The mechatronics specialty provides an educational path, that is, a roadmap, for engineering students studying within the traditional structure of most engineering colleges. Mechatronics is generally recognized worldwide as a vibrant area of study. Undergraduate and graduate programs in mechatronic engineering are now offered in many universities. Refereed journals are being published and dedicated conferences are being organized and are generally highly attended.

It should be understood that mechatronics is not just a convenient structure for investigative studies by academicians; it is a way of life in modern engineering practice. The introduction of the microprocessor in the early 1980s and the ever increasing desired performance to cost ratio revolutionized the paradigm of engineering design. The number of new products being developed at the intersection of traditional disciplines of engineering, computer science, and the natural sciences is ever increasing. New developments in these traditional disciplines are being absorbed into mechatronics design at an ever increasing pace. The ongoing information technology revolution, advances in wireless communication, smart sensors design (enabled by MEMS technology), and embedded systems engineering ensures that the engineering design paradigm will continue to evolve in the early twenty-first century.

1.2 Key Elements of Mechatronics

The study of mechatronic systems can be divided into the following areas of specialty:

1. Physical Systems Modeling
2. Sensors and Actuators
3. Signals and Systems
4. Computers and Logic Systems
5. Software and Data Acquisition

The key elements of mechatronics are illustrated in Fig. 1.1. As the field of mechatronics continues to mature, the list of relevant topics associated with the area will most certainly expand and evolve.

1.3 Historical Perspective

Attempts to construct automated mechanical systems has an interesting history. Actually, the term "automation" was not popularized until the 1940s when it was coined by the Ford Motor Company to denote a process in which a machine transferred a sub-assembly item from one station to another and then positioned the item precisely for additional assembly operations. But successful development of automated mechanical systems occurred long before then. For example, early applications of automatic control

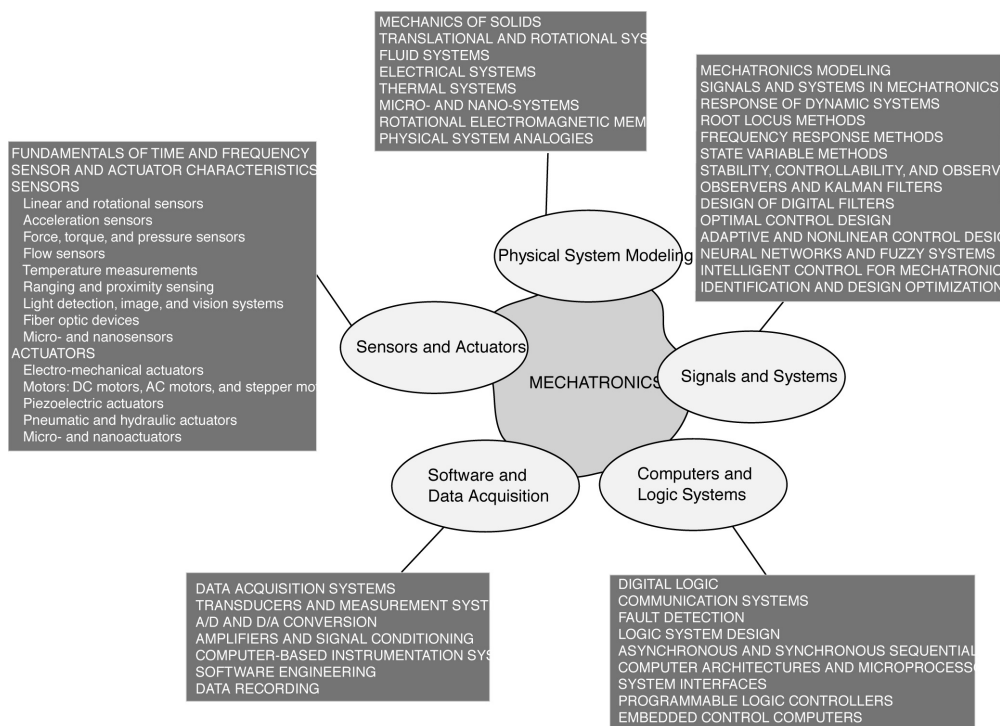



FIGURE 1.1 The key elements of mechatronics.

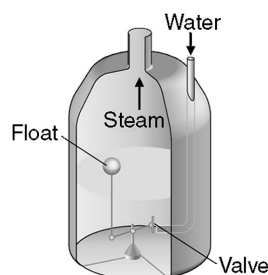



FIGURE 1.2 Water-level float regulator. (From *Modern Control Systems,* 9th ed., R. C. Dorf and R. H. Bishop, Prentice-Hall, 2001. Used with permission.)

systems appeared in Greece from 300 to 1 B.C. with the development of float regulator mechanisms [7]. Two important examples include the water clock of Ktesibios that used a float regulator, and an oil lamp devised by Philon, which also used a float regulator to maintain a constant level of fuel oil. Later, in the first century, Heron of Alexandria published a book entitled *Pneumatica* that described different types of water-level mechanisms using float regulators.

In Europe and Russia, between seventeenth and nineteenth centuries, many important devices were invented that would eventually contribute to mechatronics. Cornelis Drebbel (1572–1633) of Holland devised the temperature regulator representing one of the first feedback systems of that era. Subsequently, Dennis Papin (1647–1712) invented a pressure safety regulator for steam boilers in 1681. Papin's pressure regulator is similar to a modern-day pressure-cooker valve. The first mechanical calculating machine was invented by Pascal in 1642 [8]. The first historical feedback system claimed by Russia was developed by Polzunov in 1765 [9]. Polzunov's water-level float regulator, illustrated in Fig. 1.2, employs a float that rises and lowers in relation to the water level, thereby controlling the valve that covers the water inlet in the boiler.

Further evolution in automation was enabled by advancements in control theory traced back to the Watt flyball governor of 1769. The flyball governor, illustrated in Fig. 1.3, was used to control the speed

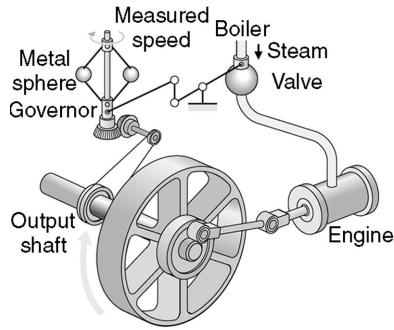



FIGURE 1.3 Watt's flyball governor. (From *Modern Control Systems,* 9th ed., R. C. Dorf and R. H. Bishop, Prentice-Hall, 2001. Used with permission.)

of a steam engine [10]. Employing a measurement of the speed of the output shaft and utilizing the motion of the flyball to control the valve, the amount of steam entering the engine is controlled. As the speed of the engine increases, the metal spheres on the governor apparatus rise and extend away from the shaft axis, thereby closing the valve. This is an example of a feedback control system where the feedback signal and the control actuation are completely coupled in the mechanical hardware.

These early successful automation developments were achieved through intuition, application of practical skills, and persistence. The next step in the evolution of automation required a *theory* of automatic control. The precursor to the numerically controlled (NC) machines for automated manufacturing (to be developed in the 1950s and 60s at MIT) appeared in the early 1800s with the invention of feed-forward control of weaving looms by Joseph Jacquard of France. In the late 1800s, the subject now known as control theory was initiated by J. C. Maxwell through analysis of the set of differential equations describing the flyball governor [11]. Maxwell investigated the effect various system parameters had on the system performance. At about the same time, Vyshnegradskii formulated a mathematical theory of regulators [12]. In the 1830s, Michael Faraday described the law of induction that would form the basis of the electric motor and the electric dynamo. Subsequently, in the late 1880s, Nikola Tesla invented the alternating-current induction motor. The basic idea of controlling a mechanical system automatically was firmly established by the end of 1800s. The evolution of automation would accelerate significantly in the twentieth century.

The development of pneumatic control elements in the 1930s matured to a point of finding applications in the process industries. However, prior to 1940, the design of control systems remained an art generally characterized by trial-and-error methods. During the 1940s, continued advances in mathematical and analytical methods solidified the notion of control engineering as an independent engineering discipline. In the United States, the development of the telephone system and electronic feedback amplifiers spurred the use of feedback by Bode, Nyquist, and Black at Bell Telephone Laboratories [13–17]. The operation of the feedback amplifiers was described in the frequency domain and the ensuing design and analysis practices are now generally classified as "classical control." During the same time period, control theory was also developing in Russia and eastern Europe. Mathematicians and applied mechanicians in the former Soviet Union dominated the field of controls and concentrated on time domain formulations and differential equation models of systems. Further developments of time domain formulations using state variable system representations occurred in the 1960s and led to design and analysis practices now generally classified as "modern control."

The World War II war effort led to further advances in the theory and practice of automatic control in an effort to design and construct automatic airplane pilots, gun-positioning systems, radar antenna control systems, and other military systems. The complexity and expected performance of these military systems necessitated an extension of the available control techniques and fostered interest in control systems and the development of new insights and methods. Frequency domain techniques continued to dominate the field of controls following World War II, with the increased use of the Laplace transform, and the use of the so-called *s*-plane methods, such as designing control systems using root locus.

On the commercial side, driven by cost savings achieved through mass production, automation of the production process was a high priority beginning in the 1940s. During the 1950s, the invention of the cam, linkages, and chain drives became the major enabling technologies for the invention of new products and high-speed precision manufacturing and assembly. Examples include textile and printing machines, paper converting machinery, and sewing machines. High-volume precision manufacturing became a reality during this period. The automated paperboard container-manufacturing machine employs a sheet-fed process wherein the paperboard is cut into a fan shape to form the tapered sidewall, and wrapped around a mandrel. The seam is then heat sealed and held until cured. Another sheet-fed source of paperboard is used to cut out the plate to form the bottom of the paperboard container, formed into a shallow dish through scoring and creasing operations in a die, and assembled to the cup shell. The lower edge of the cup shell is bent inwards over the edge of the bottom plate sidewall, and heat-sealed under high pressure to prevent leaks and provide a precisely level edge for standup. The brim is formed on the top to provide a ring-on-shell structure to provide the stiffness needed for its functionality. All of these operations are carried out while the work piece undergoes a precision transfer from one turret to another and is then ejected. The production rate of a typical machine averages over 200 cups per minute. The automated paperboard container manufacturing did not involve any non-mechanical system except an electric motor for driving the line shaft. These machines are typical of paper converting and textile machinery and represent automated systems significantly more complex than their predecessors.

The development of the microprocessor in the late 1960s led to early forms of computer control in process and product design. Examples include numerically controlled (NC) machines and aircraft control systems. Yet the manufacturing processes were still entirely mechanical in nature and the automation and control systems were implemented only as an afterthought. The launch of Sputnik and the advent of the space age provided yet another impetus to the continued development of controlled mechanical systems. Missiles and space probes necessitated the development of complex, highly accurate control systems. Furthermore, the need to minimize satellite mass (that is, to minimize the amount of fuel required for the mission) while providing accurate control encouraged advancements in the important field of optimal control. Time domain methods developed by Liapunov, Minorsky, and others, as well as the theories of optimal control developed by L. S. Pontryagin in the former Soviet Union and R. Bellman in the United States, were well matched with the increasing availability of high-speed computers and new programming languages for scientific use.

Advancements in semiconductor and integrated circuits manufacturing led to the development of a new class of products that incorporated mechanical and electronics in the system and required the two together for their functionality. The term mechatronics was introduced by Yasakawa Electric in 1969 to represent such systems. Yasakawa was granted a trademark in 1972, but after widespread usage of the term, released its trademark rights in 1982 [1–3]. Initially, mechatronics referred to systems with only mechanical systems and electrical components—no computation was involved. Examples of such systems include the automatic sliding door, vending machines, and garage door openers.

In the late 1970s, the Japan Society for the Promotion of Machine Industry (JSPMI) classified mechatronics products into four categories [1]:

1. *Class I:* Primarily mechanical products with electronics incorporated to enhance functionality. Examples include numerically controlled machine tools and variable speed drives in manufacturing machines.
2. *Class II:* Traditional mechanical systems with significantly updated internal devices incorporating electronics. The external user interfaces are unaltered. Examples include the modern sewing machine and automated manufacturing systems.
3. *Class III:* Systems that retain the functionality of the traditional mechanical system, but the internal mechanisms are replaced by electronics. An example is the digital watch.
4. *Class IV:* Products designed with mechanical and electronic technologies through synergistic integration. Examples include photocopiers, intelligent washers and dryers, rice cookers, and automatic ovens.

The enabling technologies for each mechatronic product class illustrate the progression of electromechanical products in stride with developments in control theory, computation technologies, and microprocessors. Class I products were enabled by servo technology, power electronics, and control theory. Class II products were enabled by the availability of early computational and memory devices and custom circuit design capabilities. Class III products relied heavily on the microprocessor and integrated circuits to replace mechanical systems. Finally, Class IV products marked the beginning of true mechatronic systems, through integration of mechanical systems and electronics. It was not until the 1970s with the development of the microprocessor by the Intel Corporation that integration of computational systems with mechanical systems became practical.

The divide between classical control and modern control was significantly reduced in the 1980s with the advent of "robust control" theory. It is now generally accepted that control engineering must consider both the time domain and the frequency domain approaches simultaneously in the analysis and design of control systems. Also, during the 1980s, the utilization of digital computers as integral components of control systems became routine. There are literally hundreds of thousands of digital process control computers installed worldwide [18,19]. Whatever definition of mechatronics one chooses to adopt, it is evident that modern mechatronics involves computation as the central element. In fact, the incorporation of the microprocessor to precisely modulate mechanical power and to adapt to changes in environment are the essence of modern mechatronics and smart products.

1.4 The Development of the Automobile as a Mechatronic System

The evolution of modern mechatronics can be illustrated with the example of the automobile. Until the 1960s, the radio was the only significant electronics in an automobile. All other functions were entirely mechanical or electrical, such as the starter motor and the battery charging systems. There were no "intelligent safety systems," except augmenting the bumper and structural members to protect occupants in case of accidents. Seat belts, introduced in the early 1960s, were aimed at improving occupant safety and were completely mechanically actuated. All engine systems were controlled by the driver and/or other mechanical control systems. For instance, before the introduction of sensors and microcontrollers, a mechanical distributor was used to select the specific spark plug to fire when the fuel–air mixture was compressed. The timing of the ignition was the control variable. The mechanically controlled combustion process was not optimal in terms of fuel efficiency. Modeling of the combustion process showed that, for increased fuel efficiency, there existed an optimal time when the fuel should be ignited. The timing depends on load, speed, and other measurable quantities. The electronic ignition system was one of the first mechatronic systems to be introduced in the automobile in the late 1970s. The electronic ignition system consists of a crankshaft position sensor, camshaft position sensor, airflow rate, throttle position, rate of throttle position change sensors, and a dedicated microcontroller determining the timing of the spark plug firings. Early implementations involved only a Hall effect sensor to sense the position of the rotor in the distributor accurately. Subsequent implementations eliminated the distributor completely and directly controlled the firings utilizing a microprocessor.

The Antilock Brake System (ABS) was also introduced in the late 1970s in automobiles [20]. The ABS works by sensing lockup of any of the wheels and then modulating the hydraulic pressure as needed to minimize or eliminate sliding. The Traction Control System (TCS) was introduced in automobiles in the mid-1990s. The TCS works by sensing slippage during acceleration and then modulating the power to the slipping wheel. This process ensures that the vehicle is accelerating at the maximum possible rate under given road and vehicle conditions. The Vehicle Dynamics Control (VDC) system was introduced in automobiles in the late 1990s. The VDC works similar to the TCS with the addition of a yaw rate sensor and a lateral accelerometer. The driver intention is determined by the steering wheel position and then compared with the actual direction of motion. The TCS system is then activated to control the

power to the wheels and to control the vehicle velocity and minimize the difference between the steering wheel direction and the direction of the vehicle motion [20,21]. In some cases, the ABS is used to slow down the vehicle to achieve desired control. In automobiles today, typically, 8, 16, or 32-bit CPUs are used for implementation of the various control systems. The microcontroller has onboard memory (EEPROM/EPROM), digital and analog inputs, A/D converters, pulse width modulation (PWM), timer functions, such as event counting and pulse width measurement, prioritized inputs, and in some cases digital signal processing. The 32-bit processor is used for engine management, transmission control, and airbags; the 16-bit processor is used for the ABS, TCS, VDC, instrument cluster, and air conditioning systems; the 8-bit processor is used for seat, mirror control, and window lift systems. Today, there are about 30–60 microcontrollers in a car. This is expected to increase with the drive towards developing modular systems for plug-n-ply mechatronics subsystems.

Mechatronics has become a necessity for product differentiation in automobiles. Since the basics of internal combustion engine were worked out almost a century ago, differences in the engine design among the various automobiles are no longer useful as a product differentiator. In the 1970s, the Japanese automakers succeeded in establishing a foothold in the U.S. automobile market by offering unsurpassed quality and fuel-efficient small automobiles. The quality of the vehicle was the product differentiator through the 1980s. In the 1990s, consumers came to expect quality and reliability in automobiles from all manufacturers. Today, *mechatronic features* have become the product differentiator in these traditionally mechanical systems. This is further accelerated by higher performance price ratio in electronics, market demand for innovative products with smart features, and the drive to reduce cost of manufacturing of existing products through redesign incorporating mechatronics elements. With the prospects of low single digit (2–3%) growth, automotive makers will be searching for high-tech features that will differentiate their vehicles from others [22]. The automotive electronics market in North America, now at about $20 billion, is expected to reach $28 billion by 2004 [22]. New applications of mechatronic systems in the automotive world include semi-autonomous to fully autonomous automobiles, safety enhancements, emission reduction, and other features including intelligent cruise control, and brake by wire systems eliminating the hydraulics [23]. Another significant growth area that would benefit from a mechatronics design approach is wireless networking of automobiles to ground stations and vehicle-to-vehicle communication. Telematics, which combines audio, hands-free cell phone, navigation, Internet connectivity, e-mail, and voice recognition, is perhaps the largest potential automotive growth area. In fact, the use of electronics in automobiles is expected to increase at an annual rate of 6% per year over the next five years, and the electronics functionality will double over the next five years [24].

Micro Electromechanical Systems (MEMS) is an enabling technology for the cost-effective development of sensors and actuators for mechatronics applications. Already, several MEMS devices are in use in automobiles, including sensors and actuators for airbag deployment and pressure sensors for manifold pressure measurement. Integrating MEMS devices with CMOS signal conditioning circuits on the same silicon chip is another example of development of enabling technologies that will improve mechatronic products, such as the automobile.

Millimeter wave radar technology has recently found applications in automobiles. The millimeter wave radar detects the location of objects (other vehicles) in the scenery and the distance to the obstacle and the velocity in real-time. A detailed description of a working system is given by Suzuki et al. [25]. Figure 1.4 shows an illustration of the vehicle-sensing capability with a millimeter-waver radar. This technology provides the capability to control the distance between the vehicle and an obstacle (or another vehicle) by integrating the sensor with the cruise control and ABS systems. The driver is able to set the speed and the desired distance between the cars ahead of him. The ABS system and the cruise control system are coupled together to safely achieve this remarkable capability. One logical extension of the obstacle avoidance capability is slow speed semi-autonomous driving where the vehicle maintains a constant distance from the vehicle ahead in traffic jam conditions. Fully autonomous vehicles are well within the scope of mechatronics development within the next 20 years. Supporting investigations are underway in many research centers on development of semi-autonomous cars with reactive path planning using GPS-based continuous traffic model updates and stop-and-go automation. A proposed sensing and control

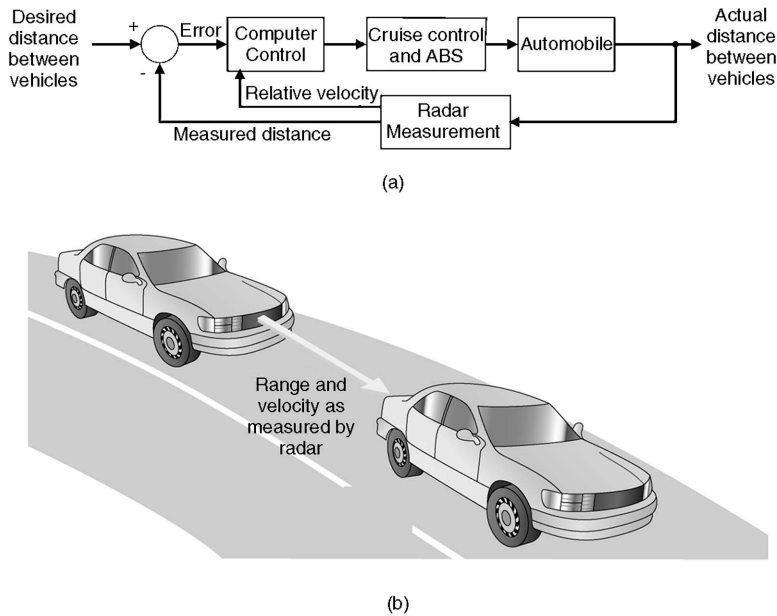



FIGURE 1.4 Using a radar to measure distance and velocity to autonomously maintain desired distance between vehicles. (Adapted from *Modern Control Systems,* 9th ed., R. C. Dorf and R. H. Bishop, Prentice-Hall, 2001. Used with permission.)

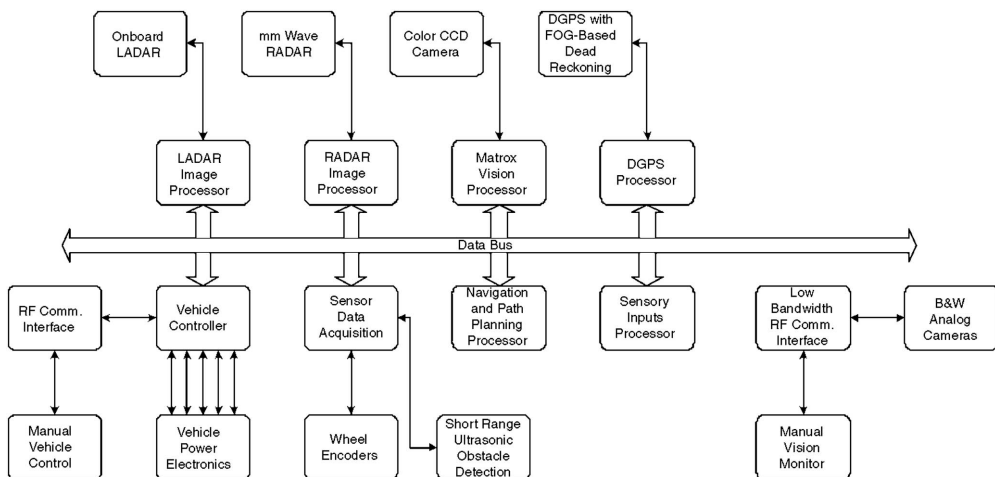



FIGURE 1.5 Autonomous vehicle system design with sensors and actuators.

system for such a vehicle, shown in Fig. 1.5, involves differential global positioning systems (DGPS), real-time image processing, and dynamic path planning [26].

Future mechatronic systems on automobiles may include a fog-free windshield based on humidity and temperature sensing and climate control, self-parallel parking, rear parking aid, lane change assistance, fluidless electronic brake-by-wire, and replacement of hydraulic systems with electromechanical servo systems. As the number of automobiles in the world increases, stricter emission standards are inevitable. Mechatronic products will in all likelihood contribute to meet the challenges in emission control and engine efficiency by providing substantial reduction in CO, NO, and HC emissions and increase in vehicle

efficiency [23]. Clearly, an automobile with 30–60 microcontrollers, up to 100 electric motors, about 200 pounds of wiring, a multitude of sensors, and thousands of lines of software code can hardly be classified as a strictly mechanical system. The automobile is being transformed into a comprehensive mechatronic system.

1.5 What is Mechatronics? And What's Next?

Mechatronics, the term coined in Japan in the 1970s, has evolved over the past 25 years and has led to a special breed of intelligent products. What is mechatronics? It is a natural stage in the evolutionary process of modern engineering design. For some engineers, mechatronics is nothing new, and, for others, it is a philosophical approach to design that serves as a guide for their activities. Certainly, mechatronics is an evolutionary process, not a revolutionary one. It is clear that an all-encompassing definition of mechatronics does not exist, but in reality, one is not needed. It is understood that mechatronics is about the synergistic integration of mechanical, electrical, and computer systems. One can understand the extent that mechatronics reaches into various disciplines by characterizing the constituent components comprising mechatronics, which include (i) physical systems modeling, (ii) sensors and actuators, (iii) signals and systems, (iv) computers and logic systems, and (v) software and data acquisition. Engineers and scientists from all walks of life and fields of study can contribute to mechatronics. As engineering and science boundaries become less well defined, more students will seek a multi-disciplinary education with a strong design component. Academia should be moving towards a curriculum, which includes coverage of mechatronic systems.

In the future, growth in mechatronic systems will be fueled by the growth in the constituent areas. Advancements in traditional disciplines fuel the growth of mechatronics systems by providing "enabling technologies." For example, the invention of the microprocessor had a profound effect on the redesign of mechanical systems and design of new mechatronics systems. We should expect continued advancements in cost-effective microprocessors and microcontrollers, sensor and actuator development enabled by advancements in applications of MEMS, adaptive control methodologies and real-time programming methods, networking and wireless technologies, mature CAE technologies for advanced system modeling, virtual prototyping, and testing. The continued rapid development in these areas will only accelerate the pace of smart product development. The Internet is a technology that, when utilized in combination with wireless technology, may also lead to new mechatronic products. While developments in automotives provide vivid examples of mechatronics development, there are numerous examples of intelligent systems in all walks of life, including smart home appliances such as dishwashers, vacuum cleaners, microwaves, and wireless network enabled devices. In the area of "human-friendly machines" (a term used by H. Kobayashi [27]), we can expect advances in robot-assisted surgery, and implantable sensors and actuators. Other areas that will benefit from mechatronic advances may include robotics, manufacturing, space technology, and transportation. The future of mechatronics is wide open.

References

1. Kyura, N. and Oho, H., "Mechatronics—an industrial perspective," *IEEE/ASME Transactions on Mechatronics,* Vol. 1, No. 1, 1996, pp. 10–15.
2. Mori, T., "Mechatronics," Yasakawa Internal Trademark Application Memo 21.131.01, July 12, 1969.
3. Harshama, F., Tomizuka, M., and Fukuda, T., "Mechatronics—What is it, why, and how?—an editorial," *IEEE/ASME Transactions on Mechatronics,* Vol. 1, No. 1, 1996, pp. 1–4.
4. Auslander, D. M. and Kempf, C. J., *Mechatronics: Mechanical System Interfacing,* Prentice-Hall, Upper Saddle River, NJ, 1996.
5. Shetty, D. and Kolk, R. A., *Mechatronic System Design,* PWS Publishing Company, Boston, MA, 1997.
6. Bolton, W., *Mechatronics: Electrical Control Systems in Mechanical and Electrical Engineering, 2nd Ed.,* Addison-Wesley Longman, Harlow, England, 1999.
7. Mayr, I. O., *The Origins of Feedback Control,* MIT Press, Cambridge, MA, 1970.

8. Tomkinson, D. and Horne, J., *Mechatronics Engineering,* McGraw-Hill, New York, 1996.
9. Popov, E. P., *The Dynamics of Automatic Control Systems;* Gostekhizdat, Moscow, 1956; Addison-Wesley, Reading, MA, 1962.
10. Dorf, R. C. and Bishop, R. H., *Modern Control Systems, 9th Ed.,* Prentice-Hall, Upper Saddle River, NJ, 2000.
11. Maxwell, J. C., "On governors," *Proc. Royal Soc. London,* 16, 1868; in *Selected Papers on Mathematical Trends in Control Theory,* Dover, New York, 1964, pp. 270–283.
12. Vyshnegradskii, I. A., "On controllers of direct action," *Izv. SPB Tekhnotog. Inst.,* 1877.
13. Bode, H. W., "Feedback—the history of an idea," in *Selected Papers on Mathematical Trends in Control Theory,* Dover, New York, 1964, pp. 106–123.
14. Black, H. S., "Inventing the Negative Feedback Amplifier," *IEEE Spectrum,* December 1977, pp. 55–60.
15. Brittain, J. E., *Turning Points in American Electrical History,* IEEE Press, New York, 1977.
16. Fagen, M. D., *A History of Engineering and Science on the Bell Systems,* Bell Telephone Laboratories, 1978.
17. Newton, G., Gould, L., and Kaiser, J., *Analytical Design of Linear Feedback Control,* John Wiley & Sons, New York, 1957.
18. Dorf, R. C. and Kusiak, A., *Handbook of Automation and Manufacturing,* John Wiley & Sons, New York, 1994.
19. Dorf, R. C., *The Encyclopedia of Robotics,* John Wiley & Sons, New York, 1988.
20. Asami, K., Nomura, Y., and Naganawa, T., "Traction Control (TRC) System for 1987 Toyota Crown, 1989," *ABS-TCS-VDC Where Will the Technology Lead Us?* J. Mack, ed., Society of Automotive Engineers, Warrendale PA, 1996.
21. Pastor, S. et al., "Brake Control System," United States Patent # 5,720,533, Feb. 24, 1998 (see http://www.uspto.gov/ for more information).
22. Jorgensen, B., "Shifting gears," Auto Electronics*, Electronic Business,* Feb. 2001.
23. Barron, M. B. and Powers, W. F., "The role of electronic controls for future automotive mechatronic systems," *IEEE/ASME Transactions on Mechatronics,* Vol. 1, No. 1, 1996, pp. 80–88.
24. Kobe, G., "Electronics: What's driving the growth?" *Automotive Industries,* August 2000.
25. Suzuki, H., Hiroshi, M. Shono, and Isaji, O., "Radar Apparatus for Detecting a Distance/Velocity," United States Patent # 5,677,695, Oct 14, 1997 (see http://www.uspto.gov/ for more information).
26. Ramasubramanian, M. K., "Mechatronics—the future of mechanical engineering-past, present, and a vision for the future," (Invited paper), *Proc. SPIE,* Vol. 4334-34, March 2001.
27. Kobayashi, H. (Guest Editorial), *IEEE/ASME Transactions on Mechatronics,* Vol. 2, No. 4, 1997, p. 217.

2

Mechatronic Design Approach

Rolf Isermann
Darmstadt University of Technology

2.1 Historical Development and Definition of Mechatronic Systems

In several technical areas the integration of products or processes and electronics can be observed. This is especially true for mechanical systems which developed since about 1980. These systems changed from electro-mechanical systems with discrete electrical and mechanical parts to integrated electronic-mechanical systems with sensors, actuators, and digital microelectronics. These integrated systems, as seen in Table 2.1, are called *mechatronic systems*, with the connection of MECHAnics and elecTRONICS.

The word “mechatronics” was probably first created by a Japanese engineer in 1969 [1], with earlier definitions given by [2] and [3]. In [4], a preliminary definition is given: “Mechatronics is the synergetic integration of mechanical engineering with electronics and intelligent computer control in the design and manufacturing of industrial products and processes” [5].

All these definitions agree that mechatronics is an *interdisciplinary field*, in which the following disciplines act together (see Fig. 2.1):

- *mechanical systems* (mechanical elements, machines, precision mechanics);
- *electronic systems* (microelectronics, power electronics, sensor and actuator technology); and
- *information technology* (systems theory, automation, software engineering, artificial intelligence).

TABLE 2.1 Historical Development of Mechanical, Electrical, and Electronic Systems

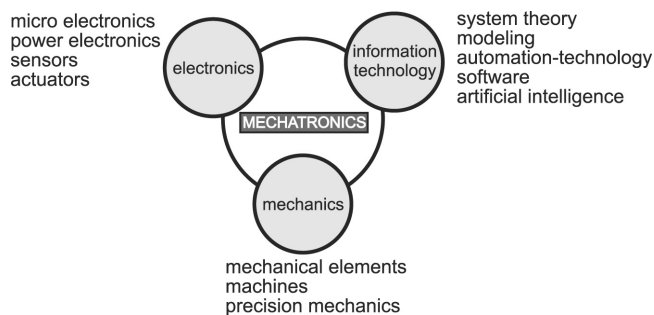
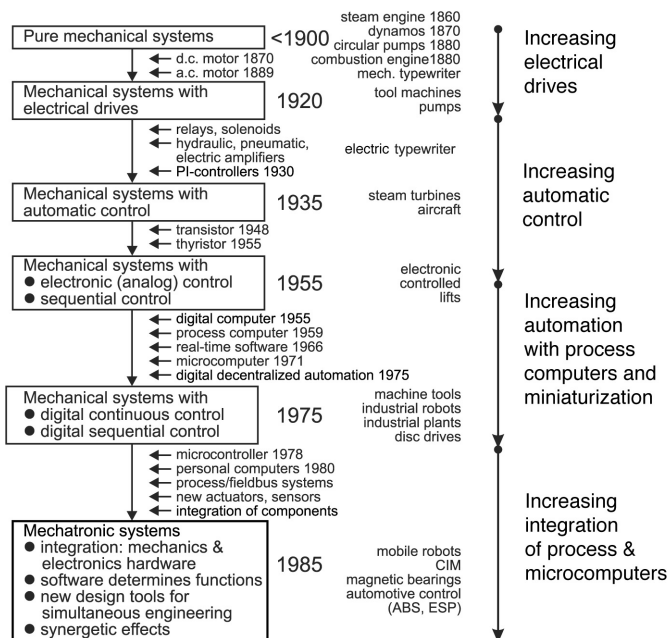



FIGURE 2.1 Mechatronics: synergetic integration of different disciplines.

Some survey contributions describe the development of mechatronics; see [5–8]. An insight into general aspects are given in the journals [4,9,10]; first conference proceedings in [11–15]; and the books [16–19].

Figure 2.2 shows a general scheme of a modern mechanical process like a power producing or a power generating machine. A primary *energy flows* into the machine and is then either directly used for the energy consumer in the case of an energy transformer, or converted into another energy form in the case of an energy converter. The form of energy can be electrical, mechanical (potential or kinetic, hydraulic, pneumatic), chemical, or thermal. Machines are mostly characterized by a continuous or periodic (repetitive) energy flow. For other mechanical processes, such as mechanical elements or precision mechanical devices, piecewise or intermittent energy flows are typical.

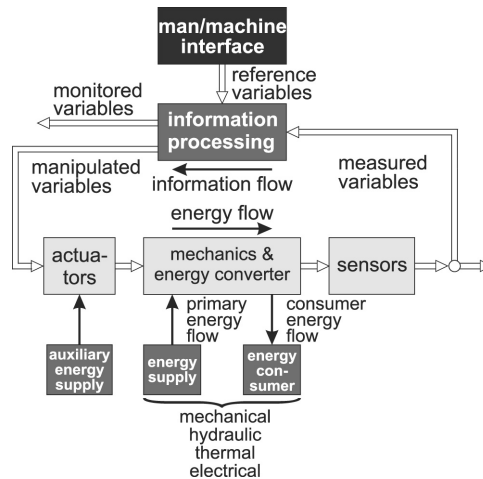



FIGURE 2.2 Mechanical process and information processing develop towards mechatronic systems.

The energy flow is generally a product of a generalized flow and a potential (effort). Information on the state of the mechanical process can be obtained by measured generalized flows (speed, volume, or mass flow) or electrical current or potentials (force, pressure, temperature, or voltage). Together with reference variables, the measured variables are the inputs for an *information flow* through the digital electronics resulting in manipulated variables for the actuators or in monitored variables on a display.

The addition and integration of feedback information flow to a feedforward energy flow in a basically mechanical system is one characteristic of many mechatronic systems. This development presently influences the design of mechanical systems. Mechatronic systems can be subdivided into:

- mechatronic systems
- mechatronic machines
- mechatronic vehicles
- precision mechatronics
- micro mechatronics

This shows that the integration with electronics comprises many classes of technical systems. In several cases, the mechanical part of the process is coupled with an electrical, thermal, thermodynamic, chemical, or information processing part. This holds especially true for energy converters as machines where, in addition to the mechanical energy, other kinds of energy appear. Therefore, *mechatronic systems in a wider sense* comprise mechanical and also non-mechanical processes. However, the mechanical part normally dominates the system.

Because an auxiliary energy is required to change the fixed properties of formerly passive mechanical systems by feedforward or feedback control, these systems are sometimes also called *active mechanical systems*.

2.2 Functions of Mechatronic Systems

Mechatronic systems permit many improved and new functions. This will be discussed by considering some examples.

Division of Functions between Mechanics and Electronics

For designing mechatronic systems, the interplay for the realization of functions in the mechanical and electronic part is crucial. Compared to pure mechanical realizations, the use of amplifiers and actuators with electrical auxiliary energy led to considerable simplifications in devices, as can be seen from watches,

electrical typewriters, and cameras. A further considerable *simplification in the mechanics* resulted from introducing microcomputers in connection with decentralized electrical drives, as can be seen from electronic typewriters, sewing machines, multi-axis handling systems, and automatic gears.

The design of lightweight constructions leads to elastic systems which are weakly damped through the material. An *electronic damping* through position, speed, or vibration sensors and electronic feedback can be realized with the additional advantage of an adjustable damping through the algorithms. Examples are elastic drive chains of vehicles with damping algorithms in the engine electronics, elastic robots, hydraulic systems, far reaching cranes, and space constructions (with, for example, flywheels).

The addition of closed loop control for position, speed, or force not only results in a precise tracking of reference variables, but also an approximate linear behavior, even though the mechanical systems show nonlinear behavior. By *omitting the constraint of linearization* on the mechanical side, the effort for construction and manufacturing may be reduced. Examples are simple mechanical pneumatic and electromechanical actuators and flow valves with electronic control.

With the aid of freely *programmable reference variable generation* the adaptation of nonlinear mechanical systems to the operator can be improved. This is already used for the driving pedal characteristics within the engine electronics for automobiles, telemanipulation of vehicles and aircraft, in development of hydraulic actuated excavators, and electric power steering.

With an increasing number of sensors, actuators, switches, and control units, the cable and electrical connections increase such that reliability, cost, weight, and the required space are major concerns. Therefore, the development of suitable bus systems, plug systems, and redundant and reconfigurable electronic systems are challenges for the designer.

Improvement of Operating Properties

By applying active feedback control, precision is obtained not only through the high mechanical precision of a passively feedforward controlled mechanical element, but by comparison of a programmed reference variable and a measured control variable. Therefore, the mechanical precision in design and manufacturing may be reduced somewhat and more simple constructions for bearings or slideways can be used. An important aspect is the compensation of a larger and time variant friction by *adaptive friction compensation* [13,20]. Also, a larger friction on cost of backlash may be intended (such as gears with pretension), because it is usually easier to compensate for friction than for backlash.

Model-based and *adaptive control* allow for a wide range of operation, compared to fixed control with unsatisfactory performance (danger of instability or sluggish behavior). A combination of robust and adaptive control allows a wide range of operation for flow-, force-, or speed-control, and for processes like engines, vehicles, or aircraft. A better control performance allows the reference variables to move closer to the constraints with an improvement in efficiencies and yields (e.g., higher temperatures, pressures for combustion engines and turbines, compressors at stalling limits, higher tensions and higher speed for paper machines and steel mills).

Addition of New Functions

Mechatronic systems allow functions to occur that could not be performed without digital electronics. First, *nonmeasurable quantities* can be calculated on the basis of measured signals and influenced by feedforward or feedback control. Examples are time-dependent variables such as slip for tyres, internal tensities, temperatures, slip angle and ground speed for steering control of vehicles, or parameters like damping, stiffness coefficients, and resistances. The *adaptation of parameters* such as damping and stiffness for oscillating systems (based on measurements of displacements or accelerations) is another example. Integrated *supervision and fault diagnosis* becomes more and more important with increasing automatic functions, increasing complexity, and higher demands on reliability and safety. Then, the triggering of redundant components, system reconfiguration, maintenance-on-request, and any kind of *teleservice* make the system more "intelligent." Table 2.2 summarizes some properties of mechatronic systems compared to conventional electro-mechanical systems.

TABLE 2.2 Properties of Conventional and Mechatronic Design Systems

	Conventional Design	Mechatronic Design
	Added components	**Integration of components (hardware)**
1	Bulky	Compact
2	Complex mechanisms	Simple mechanisms
3	Cable problems	Bus or wireless communication
4	Connected components	Autonomous units
	Simple control	**Integration by information processing (software)**
5	Stiff construction	Elastic construction with damping by electronic feedback
6	Feedforward control, linear (analog) control	Programmable feedback (nonlinear) digital control
7	Precision through narrow tolerances	Precision through measurement and feedback control
8	Nonmeasurable quantities change arbitrarily	Control of nonmeasurable estimated quantities
9	Simple monitoring	Supervision with fault diagnosis
10	Fixed abilities	Learning abilities

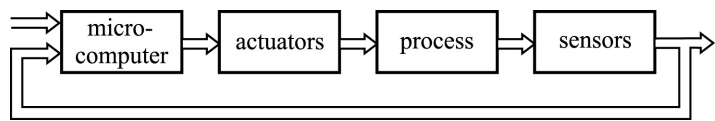



FIGURE 2.3 General scheme of a (classical) mechanical-electronic system.

2.3 Ways of Integration

Figure 2.3 shows a general scheme of a classical mechanical-electronic system. Such systems resulted from adding available sensors, actuators, and analog or digital controllers to mechanical components. The limits of this approach were given by the lack of suitable sensors and actuators, the unsatisfactory life time under rough operating conditions (acceleration, temperature, contamination), the large space requirements, the required cables, and relatively slow data processing. With increasing improvements in miniaturization, robustness, and computing power of microelectronic components, one can now put more emphasis on electronics in the design of a mechatronic system. More autonomous systems can be envisioned, such as capsuled units with touchless signal transfer or bus connections, and robust microelectronics.

The integration within a mechatronic system can be performed through the integration of components and through the integration of information processing.

Integration of Components (Hardware)

The integration of components (hardware integration) results from designing the mechatronic system as an overall system and imbedding the sensors, actuators, and microcomputers into the mechanical process, as seen in Fig. 2.4. This spatial integration may be limited to the process and sensor, or to the process and actuator. Microcomputers can be integrated with the actuator, the process or sensor, or can be arranged at several places.

Integrated sensors and microcomputers lead to *smart sensors*, and integrated actuators and microcomputers lead to *smart actuators*. For larger systems, bus connections will replace cables. Hence, there are several possibilities to build up an integrated overall system by proper integration of the hardware.

Integration of Information Processing (Software)

The integration of information processing (software integration) is mostly based on advanced control functions. Besides a basic feedforward and feedback control, an additional influence may take place through the process knowledge and corresponding online information processing, as seen in Fig. 2.4. This means a processing of available signals at higher levels, including the solution of tasks like supervision

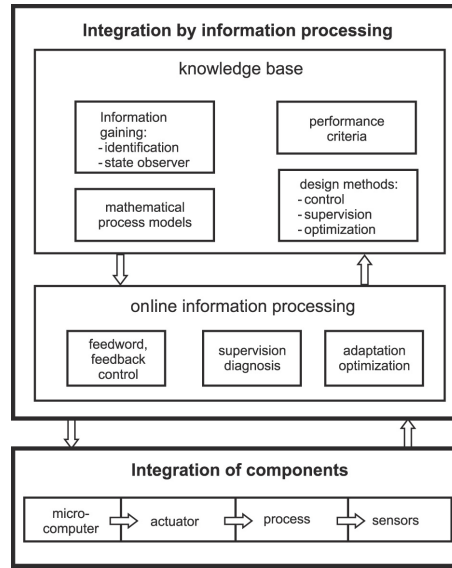



FIGURE 2.4 Ways of integration within mechatronic systems.

with fault diagnosis, optimization, and general process management. The respective problem solutions result in real-time algorithms which must be adapted to the mechanical process properties, expressed by mathematical models in the form of static characteristics, or differential equations. Therefore, a *knowledge base* is required, comprising methods for design and information gaining, process models, and performance criteria. In this way, the mechanical parts are governed in various ways through higher level information processing with intelligent properties, possibly including learning, thus forming an integration by process-adapted software.

2.4 Information Processing Systems (Basic Architecture and HW/SW Trade-offs)

The governing of mechanical systems is usually performed through actuators for the changing of positions, speeds, flows, forces, torques, and voltages. The directly measurable output quantities are frequently positions, speeds, accelerations, forces, and currents.

Multilevel Control Architecture

The information processing of *direct measurable input and output signals* can be organized in several levels, as compared in Fig. 2.5.

level 1: low level control (feedforward, feedback for damping, stabilization, linearization)
level 2: high level control (advanced feedback control strategies)
level 3: supervision, including fault diagnosis
level 4: optimization, coordination (of processes)
level 5: general process management

Recent approaches to mechatronic systems use signal processing in the lower levels, such as damping, control of motions, or simple supervision. Digital information processing, however, allows for the solution of many tasks, like adaptive control, learning control, supervision with fault diagnosis, decisions

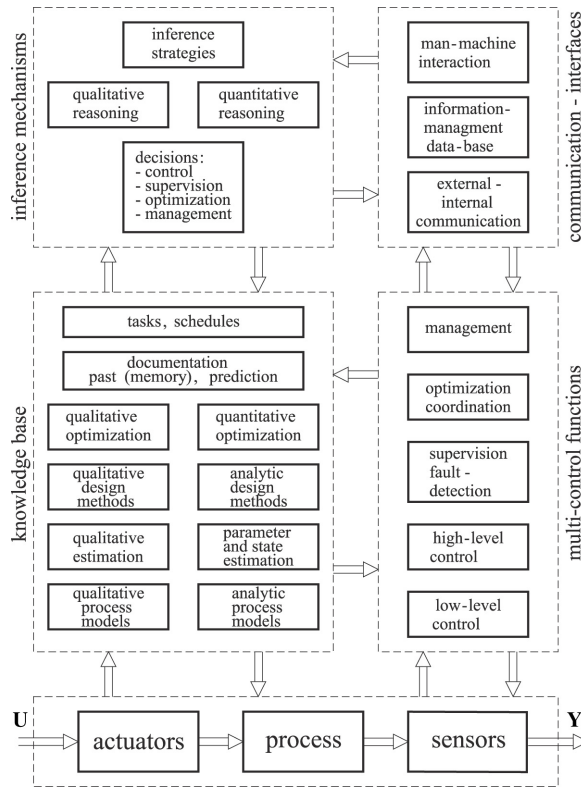



FIGURE 2.5 Advanced intelligent automatic system with multi-control levels, knowledge base, inference mechanisms, and interfaces.

for maintenance or even redundancy actions, economic optimization, and coordination. The tasks of the higher levels are sometimes summarized as "process management."

Special Signal Processing

The described methods are partially applicable for *nonmeasurable quantities* that are reconstructed from mathematical process models. In this way, it is possible to control damping ratios, material and heat stress, and slip, or to supervise quantities like resistances, capacitances, temperatures within components, or parameters of wear and contamination. This signal processing may require *special filters* to determine amplitudes or frequencies of vibrations, to determine derivated or integrated quantities, or *state variable observers*.

Model-based and Adaptive Control Systems

The information processing is, at least in the lower levels, performed by simple algorithms or software-modules under real-time conditions. These algorithms contain free adjustable parameters, which have to be adapted to the static and dynamic behavior of the process. In contrast to manual tuning by trial and error, the use of mathematical models allows precise and fast automatic adaptation.

The mathematical models can be obtained by identification and parameter estimation, which use the measured and sampled input and output signals. These methods are not restricted to linear models, but also allow for several classes of nonlinear systems. If the parameter estimation methods are combined with appropriate control algorithm design methods, adaptive control systems result. They can be used for permanent precise controller tuning or only for commissioning [20].

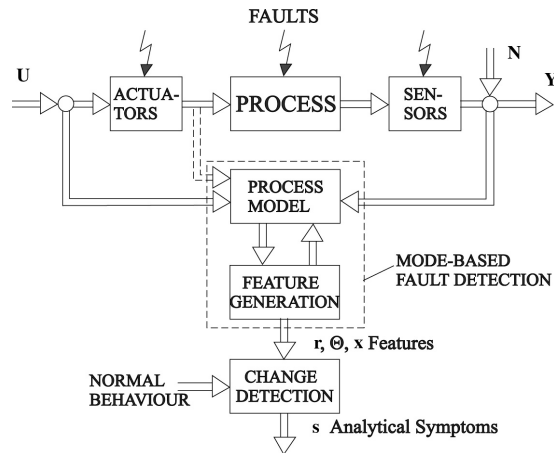



FIGURE 2.6 Scheme for a model-based fault detection.

Supervision and Fault Detection

With an increasing number of automatic functions (autonomy), including electronic components, sensors and actuators, increasing complexity, and increasing demands on reliability and safety, an integrated supervision with fault diagnosis becomes more and more important. This is a significant natural feature of an intelligent mechatronic system. Figure 2.6 shows a process influenced by faults. These faults indicate unpermitted deviations from normal states and can be generated either externally or internally. External faults can be caused by the power supply, contamination, or collision, internal faults by wear, missing lubrication, or actuator or sensor faults. The classical way for fault detection is the limit value checking of some few measurable variables. However, incipient and intermittant faults can not usually be detected, and an in-depth fault diagnosis is not possible by this simple approach. *Model-based fault detection* and *diagnosis methods* were developed in recent years, allowing for early detection of small faults with normally measured signals, also in closed loops [21]. Based on measured input signals, $\mathbf{U}(t)$, and output signals, $\mathbf{Y}(t)$, and process models, features are generated by parameter estimation, state and output observers, and parity equations, as seen in Fig. 2.6.

These residuals are then compared with the residuals for normal behavior and with change detection methods analytical symptoms are obtained. Then, a fault diagnosis is performed via methods of classification or reasoning. For further details see [22,23].

A considerable advantage is if the same process model can be used for both the (adaptive) *controller design and the fault detection.* In general, continuous time models are preferred if fault detection is based on parameter estimation or parity equations. For fault detection with state estimation or parity equations, discrete-time models can be used.

Advanced supervision and fault diagnosis is a basis for improving reliability and safety, state dependent maintenance, triggering of redundancies, and reconfiguration.

Intelligent Systems (Basic Tasks)

The information processing within mechatronic systems may range between simple control functions and intelligent control. Various definitions of intelligent control systems do exist, see [24–30]. An intelligent control system may be organized as an *online expert system*, according to Fig. 2.5, and comprises

- multi-control functions (executive functions),
- a knowledge base,
- inference mechanisms, and
- communication interfaces.

The online *control functions* are usually organized in multilevels, as already described. The *knowledge base* contains quantitative and qualitative knowledge. The quantitative part operates with analytic (mathematical) process models, parameter and state estimation methods, analytic design methods (e.g., for control and fault detection), and quantitative optimization methods. Similar modules hold for the qualitative knowledge (e.g., in the form of rules for fuzzy and soft computing). Further knowledge is the past history in the memory and the possibility to predict the behavior. Finally, tasks or schedules may be included.

The *inference mechanism* draws conclusions either by quantitative reasoning (e.g., Boolean methods) or by qualitative reasoning (e.g., possibilistic methods) and takes decisions for the executive functions.

Communication between the different modules, an information management database, and the man–machine interaction has to be organized.

Based on these functions of an online expert system, an intelligent system can be built up, with the ability "to model, reason and learn the process and its automatic functions within a given frame and to govern it towards a certain goal." Hence, intelligent mechatronic systems can be developed, ranging from "low-degree intelligent" [13], such as intelligent actuators, to "fairly intelligent systems," such as self-navigating automatic guided vehicles.

An *intelligent mechatronic system* adapts the controller to the mostly nonlinear behavior (adaptation), and stores its controller parameters in dependence on the position and load (learning), supervises all relevant elements, and performs a fault diagnosis (supervision) to request maintenance or, if a failure occurs, to request a fail safe action (decisions on actions). In the case of multiple components, supervision may help to switch off the faulty component and to perform a reconfiguration of the controlled process.

2.5 Concurrent Design Procedure for Mechatronic Systems

The design of mechatronic systems requires a systematic development and use of modern design tools.

Design Steps

Table 2.3 shows five important development steps for mechatronic systems, starting from a purely mechanical system and resulting in a fully integrated mechatronic system. Depending on the kind of mechanical system, the intensity of the single development steps is different. For precision mechanical devices, fairly integrated mechatronic systems do exist. The influence of the electronics on *mechanical elements* may be considerable, as shown by adaptive dampers, anti-lock system brakes, and automatic gears. However, complete *machines* and *vehicles* show first a mechatronic design of their elements, and then slowly a redesign of parts of the overall structure as can be observed in the development of machine tools, robots, and vehicle bodies.

Required CAD/CAE Tools

The computer aided development of mechatronic systems comprises:

1. constructive specification in the engineering development stage using CAD and CAE tools,
2. model building for obtaining static and dynamic process models,
3. transformation into computer codes for system simulation, and
4. programming and implementation of the final mechatronic software.

Some software tools are described in [31]. A broad range of CAD/CAE tools is available for 2D- and 3D-mechanical design, such as Auto CAD with a direct link to CAM (computer-aided manufacturing), and PADS, for multilayer, printed-circuit board layout. However, the state of computer-aided modeling is not as advanced. Object-oriented languages such as DYMOLA and MOBILE for modeling of large combined systems are described in [31–33]. These packages are based on specified ordinary differential

TABLE 2.3 Steps in the Design of Mechatronic Systems

	Precision Mechanics	Mechanical Elements	Machines
Pure mechanical system	large	large	large
1. Addition of sensors, actuators, microelectronics, control functions	large	large	large
2. Integration of components (hardware integration)	large	large	little
3. Integration by information processing (software integration)	large	large	little
4. Redesign of mechanical system	large	large	little
5. Creation of synergetic effects	large	little	little
Fully integrated mechatronic systems	large	little	little
Examples	Sensors actuators disc-storages cameras	Suspensions dampers clutches gears brakes	Electric drives combustion engines mach. tools robots

The size of a circle indicates the present intensity of the respective mechatronic development step: large, medium, little.

equations, algebraic equations, and discontinuities. A recent description of the state of computer-aided control system design can be found in [34]. For system simulation (and controller design), a variety of program systems exist, like ACSL, SIMPACK, MATLAB/SIMULINK, and MATRIX-X. These simulation techniques are valuable tools for design, as they allow the designer to study the interaction of components and the variations of design parameters before manufacturing. They are, in general, not suitable for real-time simulation.

Modeling Procedure

Mathematical process models for static and dynamic behavior are required for various steps in the design of mechatronic systems, such as simulation, control design, and reconstruction of variables. Two ways to obtain these models are *theoretical modeling* based on first (physical) principles and *experimental modeling* (*identification*) with measured input and output variables. A basic problem of theoretical modeling of mechatronic systems is that the components originate from different domains. There exists a well-developed domain specific knowledge for the modeling of electrical circuits, multibody mechanical systems, or hydraulic systems, and corresponding software packages. However, a computer-assisted general methodology for the modeling and simulation of components from different domains is still missing [35].

The basic principles of theoretical modeling for system with energy flow are known and can be unified for components from different domains as electrical, mechanical, and thermal (see [36–41]). The modeling methodology becomes more involved if material flows are incorporated as for fluidics, thermodynamics, and chemical processes.

A general procedure for theoretical modeling of lumped parameter processes can be sketched as follows [19].

1. Definition of flows
   - energy flow (electrical, mechanical, thermal conductance)
   - energy and material flow (fluidic, thermal transfer, thermodynamic, chemical)
2. Definition of process elements: flow diagrams
   - sources, sinks (dissipative)
   - storages, transformers, converters
3. Graphical representation of the process model
   - multi-port diagrams (terminals, flows, and potentials, or across and through variables)
   - block diagrams for signal flow
   - bond graphs for energy flow
4. Statement of equations for all process elements
   (i) Balance equations for storage (mass, energy, momentum)
   (ii) Constitutive equations for process elements (sources, transformers, converters)
   (iii) Phenomenological laws for irreversible processes (dissipative systems: sinks)
5. Interconnection equations for the process elements
   - continuity equations for parallel connections (node law)
   - compatibility equations for serial connections (closed circuit law)
6. Overall process model calculation
   - establishment of input and output variables
   - state space representation
   - input/output models (differential equations, transfer functions)

An example of steps 1–3 is shown in Fig. 2.7 for a drive-by-wire vehicle. A unified approach for processes with energy flow is known for electrical, mechanical, and hydraulic processes with incompressible fluids. Table 2.4 defines generalized through and across variables.

In these cases, the product of the through and across variable is power. This unification enabled the formulation of the standard *bond graph modeling* [39]. Also, for hydraulic processes with compressible fluids and thermal processes, these variables can be defined to result in powers, as seen in Table 2.4. However, using mass flows and heat flows is not engineering practice. If these variables are used, so-called pseudo bond graphs with special laws result, leaving the simplicity of standard bond graphs. Bond graphs lead to a high-level abstraction, have less flexibility, and need additional effort to generate simulation algorithms. Therefore, they are not the ideal tool for mechatronic systems [35]. Also, the tedious work needed to establish *block diagrams* with an early definition of causal input/output blocks is not suitable.

Development towards object-oriented modeling is on the way, where objects with terminals (cuts) are defined without assuming a causality in this basic state. Then, object diagrams are graphically represented, retaining an intuitive understanding of the original physical components [43,44]. Hence, theoretical modeling of mechatronic systems with a unified, transparent, and flexible procedure (from the basic components of different domains to simulation) are a challenge for further development. Many components show nonlinear behavior and nonlinearities (friction and backlash). For more complex process parts, multidimensional mappings (e.g., combustion engines, tire behavior) must be integrated.

For verification of theoretical models, several well-known identification methods can be used, such as correlation analysis and frequency response measurement, or Fourier- and spectral analysis. Since some parameters are unknown or changed with time, parameter estimation methods can be applied, both, for models with continuous time or discrete time (especially if the models are linear in the parameters) [42,45,46]. For the identification and approximation of nonlinear, multi-dimensional characteristics,

TABLE 2.4 Generalized Through and Across Variables for Processes with Energy Flow

System	Through Variables		Across Variables	
Electrical	Electric current	I	Electric voltage	U
Magnetic	Magnetic Flow	F	Magnetic force	Q
Mechanical				
• translation	Force	F	Velocity	w
• rotation	Torque	M	Rotational speed	ω
Hydraulic	Volume flow	$\dot{V}$	Pressure	p
Thermodynamic	Entropy flow		Temperature	T

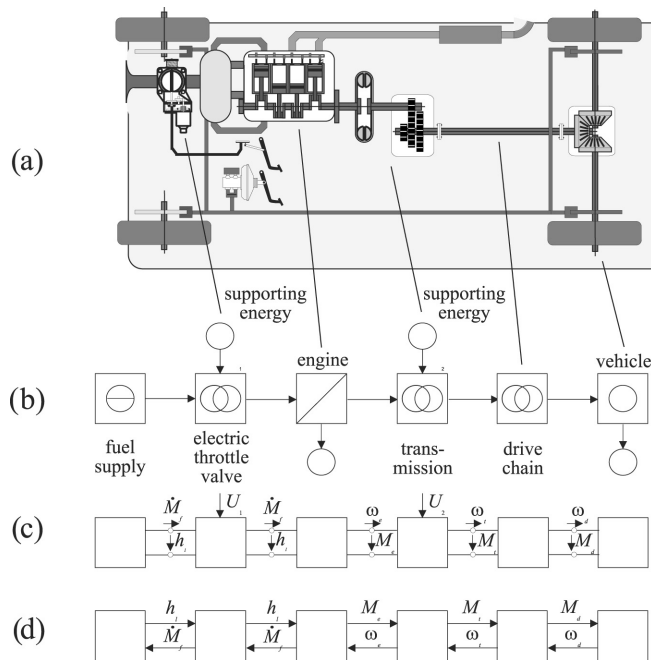



FIGURE 2.7 Different schemes for an automobile (as required for drive-by-wire-longitudinal control): (a) scheme of the components (construction map), (b) energy flow diagram (simplified), (c) multi-port diagram with flows and potentials, (d) signal flow diagram for multi-ports.

artificial neural networks (multilayer perceptrons or radial-basis-functions) can be expanded for non-linear dynamic processes [47].

Real-Time Simulation

Increasingly, real-time simulation is applied to the design of mechatronic systems. This is especially true if the process, the hardware, and the software are developed simultaneously in order to minimize iterative development cycles and to meet short time-to-market schedules. With regard to the required speed of computation *simulation methods*, it can be subdivided into

1. simulation without (hard) time limitation,
2. real-time simulation, and
3. simulation faster than real-time.

Some application examples are given in Fig. 2.8. Herewith, *real-time simulation* means that the simulation of a component is performed such that the input and output signals show the same time-dependent

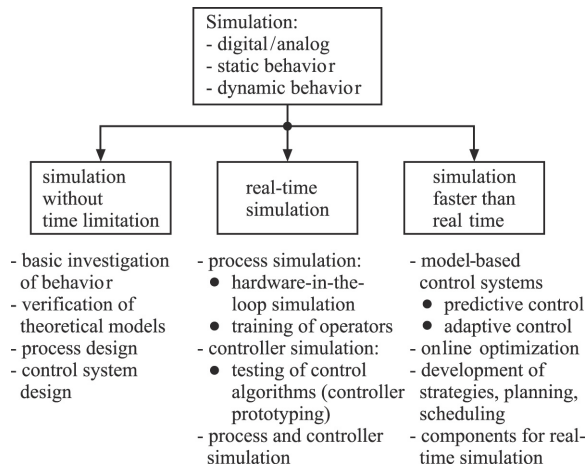



FIGURE 2.8 Classification of simulation methods with regard to speed and application examples.

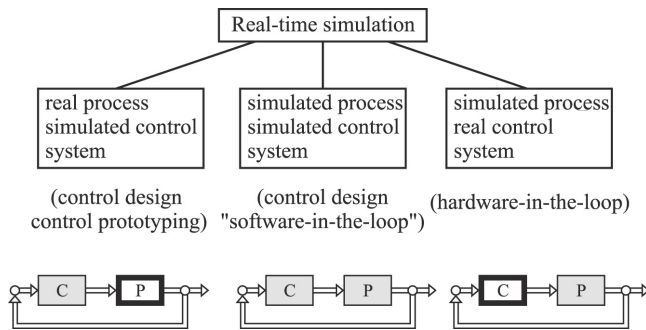



FIGURE 2.9 Classification of real-time simulation.

values as the real, dynamically operating component. This becomes a computational problem for processes which have fast dynamics compared to the required algorithms and calculation speed.

Different kinds of real-time simulation methods are shown in Fig. 2.9. The reason for the real-time requirement is mostly that one part of the investigated system is not simulated but real. Three cases can be distinguished:

1. The *real process* can be operated together with the *simulated control* by using hardware other than the final hardware. This is also called "control prototyping."
2. The *simulated process* can be operated with the *real control hardware*, which is called "hardware-in-the-loop simulation."
3. The *simulated process* is run with the *simulated control* in real time. This may be required if the final hardware is not available or if a design step before the hardware-in-the-loop simulation is considered.

Hardware-in-the-Loop Simulation

The *hardware-in-the-loop* simulation (HIL) is characterized by operating real components in connection with real-time simulated components. Usually, the control system hardware and software is the real system, as used for series production. The controlled process (consisting of actuators, physical processes, and sensors) can either comprise simulated components or real components, as seen in Fig. 2.10(a). In general, mixtures of the shown cases are realized. Frequently, some actuators are real and the process

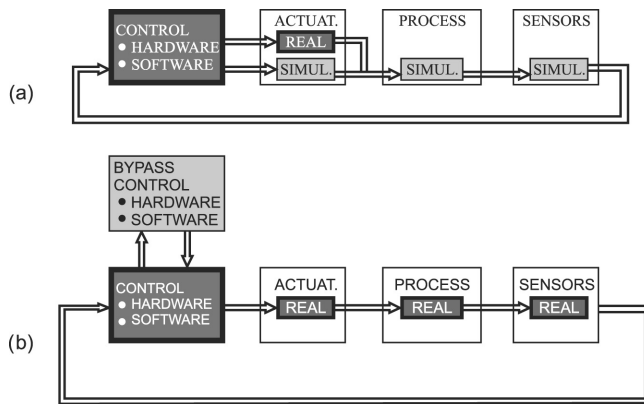



FIGURE 2.10 Real-time simulation: hybrid structures. (a) Hardware-in-the-loop simulation. (b) Control prototyping.

and the sensors are simulated. The reason is that actuators and the control hardware very often form one integrated subsystem or that actuators are difficult to model precisely and to simulate in real time. (The use of real sensors together with a simulated process may require considerable realization efforts, because the physical sensor input does not exist and must be generated artificially.) In order to change or redesign some functions of the control hardware or software, a bypass unit can be connected to the basic control hardware. Hence, hardware-in-the-loop simulators may also contain partially simulated (emulated) control functions.

The advantages of the hardware-in-the-loop simulation are generally:

- design and testing of the control hardware and software without operating a real process ("moving the process field into the laboratory");
- testing of the control hardware and software under extreme environmental conditions in the laboratory (e.g., high/low temperature, high accelerations and mechanical shocks, aggressive media, electro-magnetic compatibility);
- testing of the effects of faults and failures of actuators, sensors, and computers on the overall system;
- operating and testing of extreme and dangerous operating conditions;
- reproducible experiments, frequently repeatable;
- easy operation with different man-machine interfaces (cockpit-design and training of operators); and
- saving of cost and development time.

Control Prototyping

For the design and testing of complex control systems and their algorithms under real-time constraints, a real-time controller simulation (emulation) with hardware (e.g., off-the-shelf signal processor) other than the final series production hardware (e.g., special ASICS) may be performed. The process, the actuators, and sensors can then be real. This is called *control prototyping* (Fig. 2.10(b)). However, parts of the process or actuators may be simulated, resulting in a mixture of HIL-simulation and control prototyping. The advantages are mainly:

- early development of signal processing methods, process models, and control system structure, including algorithms with high level software and high performance off-the-shelf hardware;
- testing of signal processing and control systems, together with other design of actuators, process parts, and sensor technology, in order to create synergetic effects;

- reduction of models and algorithms to meet the requirements of cheaper mass production hardware; and
- defining the specifications for final hardware and software.

Some of the advantages of HIL-simulation also hold for control prototyping. Some references for real-time simulation are [48,49].

References

1. Kyura, N. and Oho, H., Mechatronics—an industrial perspective. *IEEE/ASME Transactions on Mechatronics,* 1(1):10–15.
2. Schweitzer, G., Mechatronik-Aufgaben und Lösungen. VDI-Berichte Nr. 787. VDI-Verlag, Düsseldorf, 1989.
3. Ovaska, S. J., Electronics and information technology in high range elevator systems. *Mechatronics,* 2(1):89–99, 1992.
4. *IEEE/ASME Transactions on Mechatronics,* 1996.
5. Harashima, F., Tomizuka, M., and Fukuda, T., Mechatronics—"What is it, why and how?" An editorial. *IEEE/ASME Transactions on Mechatronics,* 1(1):1–4, 1996.
6. Schweitzer, G., Mechatronics—a concept with examples in active magnetic bearings. *Mechatronics,* 2(1):65–74, 1992.
7. Gausemeier, J., Brexel, D., Frank, Th., and Humpert, A., Integrated product development. In *Third Conf. Mechatronics and Robotics,* Paderborn, Germany, Okt. 4–6, 1995. Teubner, Stuttgart, 1995.
8. Isermann, R., Modeling and design methodology for mechatronic systems. *IEEE/ASME Transactions on Mechatronics,* 1(1):16–28, 1996.
9. *Mechatronics: An International Journal. Aims and Scope.* Pergamon Press, Oxford, 1991.
10. *Mechatronics Systems Engineering: International Journal on Design and Application of Integrated Electromechanical Systems.* Kluwer Academic Publishers, Nethol, 1993.
11. IEE, Mechatronics: Designing intelligent machines. In *Proc. IEE-Int. Conf.* 12–13 Sep., Univ. of Cambridge, 1990.
12. Hiller, M. (ed.), *Second Conf. Mechatronics and Robotics.* September 27–29, Duisburg/Moers, Germany, 1993. Moers, IMECH, 1993.
13. Isermann, R. (ed.), Integrierte mechanisch elektroni-sche Systeme. March 2–3, Darmstadt, Germany, 1993. Fortschr.-Ber. VDI Reihe 12 Nr. 179. VDI-Verlag, Düsseldorf, 1993.
14. Lückel, J. (ed.), *Third Conf. Mechatronics and Robotics,* Paderborn, Germany, Oct. 4–6, 1995. Teubner, Stuttgart, 1995.
15. Kaynak, O., Özkan, M., Bekiroglu, N., and Tunay, I. (eds.), Recent advances in mechatronics. In *Proc. Int. Conf. Recent Advances in Mechatronics,* August 14–16, 1995, Istanbul, Turkey.
16. Kitaura, K., Industrial mechatronics. New East Business Ltd., in Japanese, 1991.
17. Bradley, D. A., Dawson, D., Burd, D., and Loader, A. J., *Mechatronics-Electronics in Products and Processes.* Chapman and Hall, London, 1991.
18. McConaill, P. A., Drews, P., and Robrock, K. H., *Mechatronics and Robotics I.* IOS-Press, Amsterdam, 1991.
19. Isermann, R., *Mechatronische Systeme.* Springer, Berlin, 1999.
20. Isermann, R., Lachmann, K. H., and Matko, D., *Adaptive Control Systems,* Prentice-Hall, London, 1992.
21. Isermann, R., Supervision, fault detection and fault diagnosis methods—advanced methods and applications. In *Proc. XIV IMEKO World Congress,* Vol. 1, pp. 1–28, Tampere, Finland, 1997.
22. Isermann, R., Supervision, fault detection and fault diagnosis methods—an introduction, special section on supervision, fault detection and diagnosis. *Control Engineering Practice,* 5(5):639–652, 1997.
23. Isermann, R. (ed.), Special section on supervision, fault detection and diagnosis. *Control Engineering Practice,* 5(5):1997.

24. Saridis, G. N., *Self Organizing Control of Stochastic Systems.* Marcel Dekker, New York, 1977.
25. Saridis, G. N. and Valavanis, K. P., Analytical design of intelligent machines. *Automatica,* 24:123–133, 1988.
26. Åström, K. J., Intelligent control. In *Proc. European Control Conf.,* Grenoble, 1991.
27. White, D. A. and Sofge, D. A. (eds.), *Handbook of Intelligent Control.* Van Norstrad, Reinhold, New York, 1992.
28. Antaklis, P., Defining intelligent control. *IEEE Control Systems,* Vol. June: 4–66, 1994.
29. Gupta, M. M. and Sinha, N. K., *Intelligent Control Systems.* IEEE-Press, New York, 1996.
30. Harris, C. J. (ed.), *Advances in Intelligent Control.* Taylor & Francis, London, 1994.
31. Otter, M. and Gruebel, G., Direct physical modeling and automatic code generation for mechatronics simulation. In *Proc. 2nd Conf. Mechatronics and Robotics,* Duisburg, Sep. 27–29, IMECH, Moers, 1993.
32. Elmquist, H., Object-oriented modeling and automatic formula manipulation in Dymola, Scandin. Simul. Society SIMS, June, Kongsberg, 1993.
33. Hiller, M., Modelling, simulation and control design for large and heavy manipulators. In *Proc. Int. Conf. Recent Advances in Mechatronics.* 1:78–85, Istanbul, Turkey, 1995.
34. James, J., Cellier, F., Pang, G., Gray, J., and Mattson, S. E., The state of computer-aided control system design (CACSD). *IEEE Transactions on Control Systems,* Special Issue, April 6–7 (1995).
35. Otter, M. and Elmqvist, H., Energy flow modeling of mechatronic systems via object diagrams. In *Proc. 2nd MATHMOD,* Vienna, 705–710, 1997.
36. Paynter, H. M., *Analysis and Design of Engineering Systems.* MIT Press, Cambridge, 1961.
37. MacFarlane, A. G. J., *Engineering Systems Analysis.* G. G. Harrop, Cambridge, 1964.
38. Wellstead, P. E., *Introduction to Physical System Modelling.* Academic Press, London, 1979.
39. Karnopp, D. C., Margolis, D. L., and Rosenberg, R. C., *System Dynamics. A Unified Approach.* J. Wiley, New York, 1990.
40. Cellier, F. E., *Continuous System Modelling.* Springer, Berlin, 1991.
41. Gawtrop, F. E. and Smith, L., *Metamodelling: Bond Graphs and Dynamic Systems.* Prentice-Hall, London, 1996.
42. Eykhoff, P., *System Identification.* John Wiley & Sons, London, 1974.
43. Elmqvist, H., A structured model language for large continuous systems. Ph.D. Dissertation, Report CODEN: LUTFD2/(TFRT-1015) Dept. of Aut. Control, Lund Institute of Technology, Sweden, 1978.
44. Elmqvist, H. and Mattson, S. E., Simulator for dynamical systems using graphics and equations for modeling. *IEEE Control Systems Magazine,* 9(1):53–58, 1989.
45. Isermann, R., *Identifikation dynamischer Systeme.* 2nd Ed., Vol. 1 and 2. Springer, Berlin, 1992.
46. Ljung, L., *System Identification: Theory for the User.* Prentice-Hall, Englewood Cliffs, NJ, 1987.
47. Isermann, R., Ernst, S., and Nelles, O., Identification with dynamic neural networks—architectures, comparisons, applications—Plenary. In *Proc. IFAC Symp. System Identification (SYSID'97),* Vol. 3, pp. 997–1022, Fukuoka, Japan, 1997.
48. Hanselmann, H., Hardware-in-the-loop simulation as a standard approach for development, customization, and production test, SAE 930207, 1993.
49. Isermann, R., Schaffnit, J., and Sinsel, S., Hardware-in-the-loop simulation for the design and testing of engine control systems. *Control Engineering Practice,* 7(7):643–653, 1999.

3

System Interfacing, Instrumentation, and Control Systems

Rick Homkes
Purdue University

3.1 Introduction

The purpose of this chapter is to introduce a number of topics dealing with a mechatronic system. This starts with an overview of mechatronic systems and a look at the input and output signals of a mechatronic system. The special features of microprocessor input and output are next. Software, an often-neglected portion of a mechatronic system, is briefly covered with an emphasis on software engineering concepts. The chapter concludes with a short discussion of testing and instrumentation.

The Mechatronic System

Figure 3.1 shows a typical mechatronic system with mechanical, electrical, and computer components. The process of system data acquisition begins with the measurement of a physical value by a sensor. The sensor is able to generate some form of signal, generally an analog signal in the form of a voltage level or waveform. This analog signal is sent to an analog-to-digital converter (ADC). Commonly using a process of successive approximation, the ADC maps the analog input signal to a digital output. This digital value is composed of a set of binary values called bits (often represented by 0s and 1s). The set of bits represents a decimal or hexadecimal number that can be used by the microcontroller. The microcontroller consists of a microprocessor plus memory and other attached devices. The program in the microprocessor uses this digital value along with other inputs and preloaded values called calibrations to determine output commands. Like the input to the microprocessor, these outputs are in digital form and can be represented by a set of bits. A digital-to-analog converter (DAC) is then often used to convert the digital value into an analog signal. The analog signal is used by an actuator to control a physical device or affect the physical environment. The sensor then takes new measurements and the process repeated, thus completing a feedback control loop. Timing for this entire operation is synchronized by the use of a clock.

A Home/Office Example

An example of a mechatronic system is the common heating/cooling system for homes and offices. Simple systems use a bimetal thermostat with contact points controlling a mercury switch that turns on and off the furnace or air conditioner. A modern environmental control system uses these same basic components along with other components and computer program control. A temperature sensor monitors the physical environment and produces a voltage level as demonstrated in Fig. 3.2 (though generally not nearly such a smooth function). After conversion by the ADC, the microcontroller uses the digitized temperature

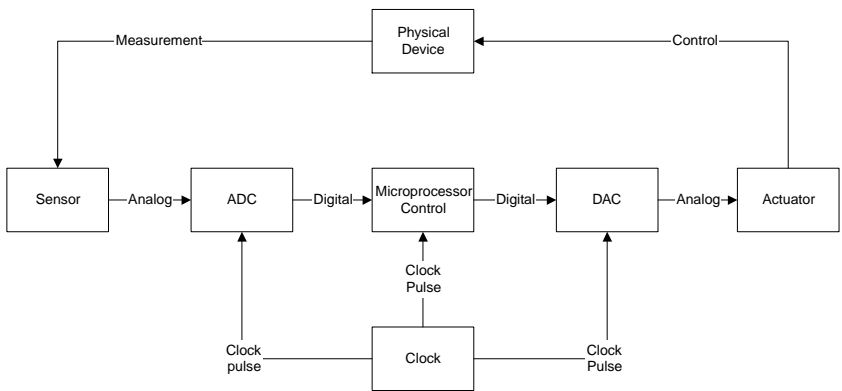



FIGURE 3.1 Microprocessor control system.

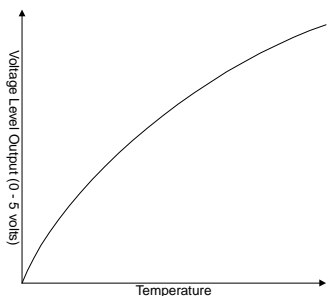



FIGURE 3.2 Voltage levels.

data along with a 24-hour clock and the user requested temperatures to produce a digital control signal. This signal directs the actuator, usually a simple electrical switch in this example. The switch, in turn, controls a motor to turn the heating or cooling unit on or off. New measurements are then taken and the cycle is repeated. While not a mechatronic product on the order of a camcorder, it is a mechatronic system because of its combination of mechanical, electrical, and computer components. This system may also incorporate some additional features. If the temperature being sensed is quite high, say 80°C, it is possible that a fire exists. It is then not a good idea to turn on the blower fan and feed the fire more oxygen. Instead the system should set off an alarm or use a data communication device to alert the fire department. Because of this type of computer control, the system is "smart," at least relative to the older mercury-switch controlled systems.

An Automotive Example

A second example is the Antilock Braking System (ABS) found in many vehicles. The entire purpose of this type of system is to prevent a wheel from locking up and thus having the driver loose directional control of the vehicle due to skidding. In this case, sensors attached to each wheel determine the rotational speed of the wheels. These data, probably in a waveform or time-varied electrical voltage, is sent to the microcontroller along with the data from sensors reporting inputs such as brake pedal position, vehicle speed, and yaw. After conversion by the ADC or input capture routine into a digital value, the program in the microprocessor then determines the necessary action. This is where the aspect of human computer interface (HCI) or human machine interface (HMI) comes into play by taking account of the "feel" of the system to the user. System calibration can adjust the response to the driver while, of course, stopping the vehicle by controlling the brakes with the actuators. There are two important things to note in this example. The first is that, in the end, the vehicle is being stopped because of hydraulic forces pressing the brake pad against a drum or rotor—a purely mechanical function. The other is that the ABS, while an "intelligent product," is not a stand-alone device. It is part of a larger system, the vehicle, with multiple microcontrollers working together through the data network of the vehicle.

3.2 Input Signals of a Mechatronic System

Transducer/Sensor Input

All inputs to mechatronic systems come from either some form of sensory apparatus or communications from other systems. Sensors were first introduced in the previous section and will be discussed in much more depth in Chapter 19. Transducers, devices that convert energy from one form to another, are often used synonymously with sensors. Transducers and their properties will be explained fully in Chapter 45. Sensors can be divided into two general classifications, active or passive. Active sensors emit a signal in order to estimate an attribute of the environment or device being measured. Passive sensors do not. A military example of this difference would be a strike aircraft "painting" a target using either active laser radar (LADAR) or a passive forward looking infrared (FLIR) sensor.

As stated in the Introduction section, the output of a sensor is usually an analog signal. The simplest type of analog signal is a voltage level with a direct (though not necessarily linear) correlation to the input condition. A second type is a pulse width modulated (PWM) signal, which will be explained further in a later section of this chapter when discussing microcontroller outputs. A third type is a waveform, as shown in Fig. 3.3. This type of signal is modulated either in its amplitude (Fig. 3.4) or its frequency (Fig. 3.5) or, in some cases, both. These changes reflect the changes in the condition being monitored.

There are sensors that do not produce an analog signal. Some of these sensors produce a square wave as in Fig. 3.6 that is input to the microcontroller using the EIA 232 communications standard. The square wave represents the binary values of 0 and 1. In this case the ADC is probably on-board the sensor itself, adding to the cost of the sensor. Some sensors/recorders can even create mail or TCP/IP packets as output. An example of this type of unit is the MV100 MobileCorder from Yokogawa Corporation of America.

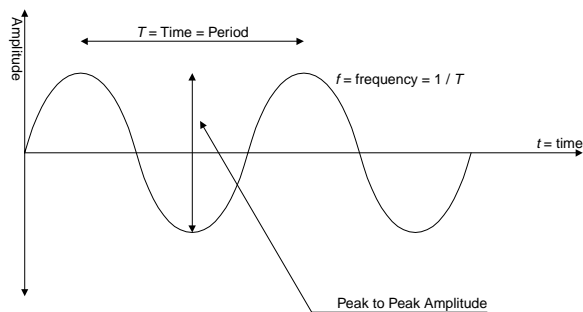



FIGURE 3.3 Sine wave.

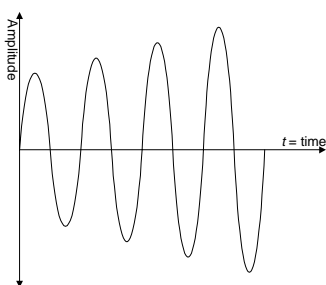



FIGURE 3.4 Amplitude modulation.

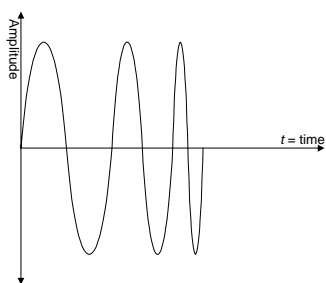



FIGURE 3.5 Frequency modulation.

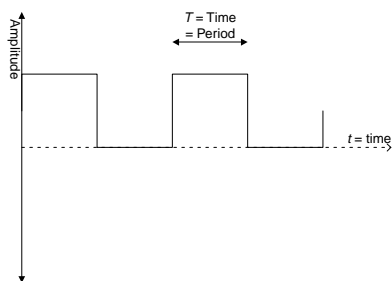



FIGURE 3.6 Square wave.

Analog-to-Digital Converters

The ADC can basically be typed by two parameters: the analog input range and the digital output range. As an example, consider an ADC that is converting a voltage level ranging 0–12 V into a single byte of 8 bits. In this example, each binary count increment reflects an increase in analog voltage of 1/256 of the maximum 12 V. There is an unusual twist to this conversion, however. Since a zero value represents 0 V, and a 128 value represents half of the maximum value, 6 V in this example, the maximum decimal value of 255 represents 255/256 of the maximum voltage value, or 11.953125 V. A table of the equivalent values is shown below:

Binary	Decimal	Voltage
0000 0000	0	0.0
0000 0001	1	0.00390625
1000 0000	128	6.0
1111 1111	255	11.953125

An ADC that is implemented in the Motorola HC12 microcontroller produces 10 bits. While not fitting so nicely into a single byte of data, this 10-bit ADC does give additional resolution. Using an input range from 0 to 5 V, the decimal resolution per least significant bit is 4.88 mV. If the ADC had 8 bits of output, the resolution per bit would be 19.5 mV, a fourfold difference. Larger voltages, e.g., from 0 to 12 V, can be scaled with a voltage divider to fit the 0–5 V range. Smaller voltages can be amplified to span the entire range. A process known as successive approximation (using the Successive Approximation Register or SAR in the Motorola chip) is used to determine the correct digital value.

3.3 Output Signals of a Mechatronic System

Digital-to-Analog Converters

The output command from the microcontroller is a binary value in bit, byte (8 bits), or word (16 bits) form. This digital signal is converted to analog using a digital-to-analog converter, or DAC. Let us examine converting an 8-bit value into a voltage level between 0 and 12 V. The most significant bit in the binary value to be converted (decimal 128) creates an analog value equal to half of the maximum output, or 6 V. The next digit produces an additional one fourth, or 3 V, the next an additional one eighth, and so forth. The sum of all these weighted output values represents the appropriate analog voltage. As was mentioned in a previous section, the maximum voltage value in the range is not obtainable, as the largest value generated is 255/256 of 12 V, or 11.953125 V. The smoothness of the signal representation depends on the number of bits accepted by the DAC and the range of the output required. Figure 3.7 demonstrates a simplified step function using a one-byte binary input and 12-V analog output.

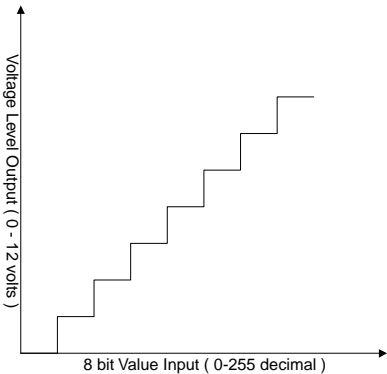



FIGURE 3.7 DAC stepped output.

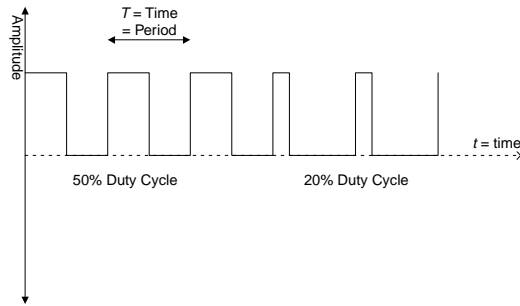



FIGURE 3.8 Pulse width modulation.

Actuator Output

Like sensors, actuators were first introduced in a previous section and will be described in detail in a later chapter of this handbook. The three common actuators that this section will review are switches, solenoids, and motors. Switches are simple state devices that control some activity, like turning on and off the furnace in a house. Types of switches include relays and solid-state devices. Solid-state devices include diodes, thyristors, bipolar transistors, field-effect transistors (FETs), and metal-oxide field-effect transistors (MOSFETs). A switch can also be used with a sensor, thus turning on or off the entire sensor, or a particular feature of a sensor.

Solenoids are devices containing a movable iron core that is activated by a current flow. The movement of this core can then control some form of hydraulic or pneumatic flow. Applications are many, including braking systems and industrial production of fluids. More information on solenoid actuators can be found in a later chapter. Motors are the last type of actuator that will be summarized here. There are three main types: direct current (DC), alternating current (AC), and stepper motors. DC motors may be controlled by a fixed DC voltage or by pulse width modulation (PWM). In a PWM signal, such as shown in Fig. 3.8, a voltage is alternately turned on and off while changing (modulating) the width of the on-time signal, or duty cycle. AC motors are generally cheaper than DC motors, but require variable frequency drive to control the rotational speed. Stepper motors move by rotating a certain number of degrees in response to an input pulse.

3.4 Signal Conditioning

Signal conditioning is the modification of a signal to make it more useful to a system. Two important types of signal conditioning are, of course, the conversion between analog and digital, as described in the previous two sections. Other types of signal conditioning are briefly covered below, with a full coverage reserved for Chapters 46 and 47.

Sampling Rate

The rate at which data samples are taken obviously affects the speed at which the mechatronic system can detect a change in situation. There are several things to consider, however. For example, the response of a sensor may be limited in time or range. There is also the time required to convert the signal into a form usable by the microprocessor, the A to D conversion time. A third is the frequency of the signal being sampled. For voice digitalization, there is a very well-known sampling rate of 8000 samples per second. This is a result of the Nyquist theorem, which states that the sampling rate, to be accurate, must be at least twice the maximum frequency being measured. The 8000 samples per second rate thus works well for converting human voice over an analog telephone system where the highest frequency is approximately 3400 Hz. Lastly, the clock speed of the microprocessor must also be considered. If the ADC and DAC are

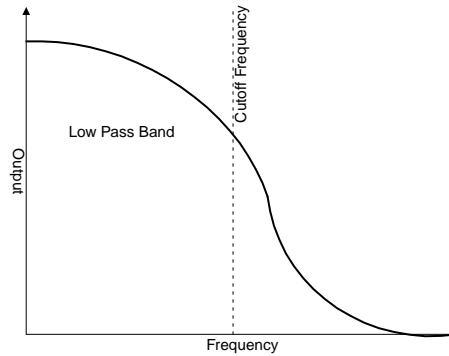



FIGURE 3.9 Low-pass filter.

on the same board as the microprocessor, they will often share a common clock. The microprocessor clock, however, may be too fast for the ADC and DAC. In this case, a prescaler is used to divide the clock frequency to a level usable by the ADC and DAC.

Filtering

Filtering is the attenuation (lessening) of certain frequencies from a signal. This process can remove noise from a signal and condition the line for better data transmission. Filters can be divided into analog and digital types, the analog filters being further divided into passive and active types. Analog passive filters use resistors, capacitors, and inductors. Analog active filters typically use operational amplifiers with resistors and capacitors. Digital filters may be implemented with software and/or hardware. The software component gives digital filters the feature of being easier to change. Digital filters are explained fully in Chapter 29.

Filters may also be differentiated by the type of frequencies they affect.

1. Low-pass filters allow lower set of frequencies to pass through, while high frequencies are attenuated. A simplistic example of this is shown in Fig. 3.9.
2. High-pass filters, the opposite of low-pass, filter a lower frequency band while allowing higher frequencies to pass.
3. Band-pass filters allow a particular range of frequencies to pass; all others are attenuated.
4. Band-stop filters stop a particular range of frequencies while all others are allowed to pass.

There are many types and applications of filters. For example, William Ribbens in his book *Understanding Automotive Electronics* (Newnes 1998) described a software low-pass filter (sometimes also called a lag filter) that averages the last 60 fuel tank level samples taken at 1 s intervals. The filtered data are then displayed on the vehicle instrument cluster. This type of filtering reduces large and quick fluctuations in the fuel gauge due to sloshing in the tank, and thus displays a more accurate value.

Data Acquisition Boards

There is a special type of board that plugs into a slot in a desktop personal computer that can be used for many of the tasks above. It is called a data acquisition board, or DAQ board. This type of board can generate analog input and multiplex multiple input signals onto a single bus for transmission to the PC. It can also come with signal conditioning hardware/software and an ADC. Some units have direct memory access (DMA), where the device writes the data directly into computer memory without using the microprocessor. While desktop PCs are not usually considered as part of a mechatronic system, the DAQ board can be very useful for instrumentation.

3.5 Microprocessor Control

PID Control

A closed loop control system is one that determines a difference in the desired and actual condition (the error) and creates a correction control command to remove this error. PID control demonstrates three ways of looking at this error and correcting it. The first way is the P of PID, the proportional term. This term represents the control action made by the microcontroller in proportion to the error. In other words, the bigger the error, the bigger the correction. The I in PID is for the integral of the error over time. The integral term produces a correction that considers the time the error has been present. Stated in other words, the longer the error continues, the bigger the correction. Lastly, the D in PID stands for derivative. In the derivative term, the corrective action is related to the derivative or change of the error with respect to time. Stated in other words, the faster the error is changing, the bigger the correction. Control systems can use P, PI, PD, or PID in creating corrective actions. The problem generally is "tuning" the system by selecting the proper values in the terms. For more information on control design, see Chapter 31.

Programmable Logic Controllers

Any discussion of control systems and microprocessor control should start with the first type of "mechatronic" control, the programmable logic controller or PLC. A PLC is a simpler, more rugged microcontroller designed for environments like a factory floor. Input is usually from switches such as push buttons controlled by machine operators or position sensors. Timers can also be programmed in the PLC to run a particular process for a set amount of time. Outputs include lamps, solenoid valves, and motors, with the input–output interfacing done within the controller. A simple programming language used with a PLC is called ladder logic or ladder programming. Ladder logic is a graphical language showing logic as a combination of series (and's) and parallel (or's) blocks. Additional information can be found in Chapter 43 and in the book *Programmable Logic Controllers* by W. Bolton (Newnes 1996).

Microprocessors

A full explanation of a microprocessor is found in section 5.8. For this discussion of microprocessors and control, we need only know a few of the component parts of computer architecture. RAM, or random access memory, is the set of memory locations the computer uses for fast temporary storage. The radio station presets selected by the driver (or passenger) in the car radio are stored in RAM. A small electrical current maintains these stored frequencies, so disconnection of the radio from the battery will result in their loss. ROM, or read only memory, is the static memory that contains the program to run the microcontroller. Thus the radio's embedded program will not be lost when the battery is disconnected. There are several types of ROM, including erasable programmable ROM (EPROM), electrically erasable programmable ROM (EEPROM), and flash memory (a newer type of EEPROM). These types will be explained later in this handbook. There are also special memory areas in a microprocessor called registers. Registers are very fast memory locations that temporarily store the address of the program instruction being executed, intermediate values needed to complete a calculation, data needed for comparison, and data that need to be input or output. Addresses and data are moved from one point to another in RAM, ROM, and registers using a bus, a set of lines transmitting data multiple bits simultaneously.

3.6 Microprocessor Numerical Control

Fixed-Point Mathematics

The microprocessors in an embedded controller are generally quite small in comparison to a personal computer or computer workstation. Adding processing power in the form of a floating-point processor and additional RAM or ROM is not always an option. This means that sometimes the complex mathematical

functions needed in a control system are not available. However, sometimes the values being sensed and computed, though real numbers, are of a reasonable range. Because of this situation there exists a special type of arithmetic whereby microcontrollers use integers in place of floating-point numbers to compute non-whole number (pseudo real) values.

There are several forms of fixed-point mathematics currently in use. The simplest form is based upon powers of 2, just like normal integers in binary. However, a virtual binary point is inserted into the integer to allow an approximation of real values to be stored as integers. A standard 8-bit unsigned integer is shown below along with its equivalent decimal value.

$$0001\ \ 0100 = (1 * 2^4) + (1 * 2^2) = (1 * 16) + (1 * 4) = 20$$

Suppose a virtual binary point is inserted between the two nibbles in the byte. There are now four bits left of the binary point with the standard positive powers of 2, and 4 bits right of the binary point with negative powers of 2. The same number now represents a real number in decimal.

$$0001\ \ 0100 = (1 * 2^0) + (1 * 2^{-2}) = (1 * 1) + (1 * 0.25) = 1.25$$

Obviously this method has shortcomings. The resolution of any fixed point number is limited to the power of 2 attached to the least significant bit on the right of the number, in this case 2^{-4} or 1/16 or 0.0625. Rounding is sometimes necessary. There is also a tradeoff in complexity, as the position of this virtual binary point must constantly be maintained when performing calculations. The savings in memory usage and processing time, however, often overcome these tradeoffs; so fixed-point mathematics can be very useful.

Calibrations

The area of calibrating a system can sometimes take on an importance not foreseen when designing a mechatronic system. The use of calibrations, numerical and logical values kept in EEPROM or ROM, allow flexibility in system tuning and implementation. For example, if different microprocessor crystal speeds may be used in a mechatronic system, but real-time values are needed, a stored calibration constant of clock cycles per microsecond will allow this calculation to be affected. Thus, calibrations are often used as a gain, the value multiplied by some input in order to produce a scaled output.

Also, as mentioned above, calibrations are often used in the testing of a mechatronic system in order to change the “feel” of the product. A transmission control unit can use a set of calibrations on engine RPM, engine load, and vehicle speed to determine when to shift gears. This is often done with hysteresis, as the shift points moving from second gear to third gear as from third gear to second gear may differ.

3.7 Microprocessor Input–Output Control

Polling and Interrupts

There are two basic methods for the microprocessor to control input and output. These are polling and interrupts. Polling is just that, the microprocessor periodically checking various peripheral devices to determine if input or output is waiting. If a peripheral device has some input or output that should be processed, a flag will be set. The problem is that a lot of processing time is wasted checking for inputs when they are not changing.

Servicing an interrupt is an alternative method to control inputs and outputs. In this method, a register in the microprocessor must have set an interrupt enable (IE) bit for a particular peripheral device. When an interrupt is initiated by the peripheral, a flag is set for the microprocessor. The interrupt request (IRQ) line will go active, and the microprocessor will service the interrupt. Servicing an interrupt means that the normal processing of the microprocessor is halted (i.e., interrupted) while the input/output is completed. In order to resume normal processing, the microprocessor needs to store the contents of its registers before the interrupt is serviced. This process includes saving all active register contents to a stack, a part

of RAM designated for this purpose, in a process known as a push. After a push, the microprocessor can then load the address of the Interrupt Service Routine and complete the input/output. When that portion of code is complete, the contents of the stack are reloaded to the registers in an operation known as a Pop (or Pull) and normal processing resumes.

Input and Output Transmission

Once the input or output is ready for transmission, there are several modes that can be used. First, data can be moved in either parallel or serial mode. Parallel mode means that multiple bits (e.g., 16 bits) move in parallel down a multiple pathway or bus from source to destination. Serial mode means that the bits move one at a time, in a series, down a single pathway. Parallel mode traffic is faster in that multiple bits are moving together, but the number of pathways is a limiting factor. For this reason parallel mode is usually used for components located close to one another while serial transmission is used if any distance is involved.

Serial data transmission can also be differentiated by being asynchronous or synchronous. Asynchronous data transmission uses separate clocks between the sender and receiver of data. Since these clocks are not synchronized, additional bits called start and stop bits are required to designate the boundaries of the bytes being sent. Synchronous data transmission uses a common or synchronized timing source. Start and stop bits are thus not needed, and overall throughput is increased.

A third way of differentiating data transmission is by direction. A simplex line is a one direction only pathway. Data from a sensor to the microcontroller may use simplex mode. Half-duplex mode allows two-way traffic, but only one direction at a time. This requires a form of flow control to avoid data transmission errors. Full-duplex mode allows two-way simultaneous transmission of data.

The agreement between sending and receiving units regarding the parameters of data transmission (including transmission speed) is known as handshaking.

HC12 Microcontroller Input–Output Subsystems

There are four input–output subsystems on the Motorola HC12 microcontroller that can be used to exemplify the data transmission section above.

The serial communications interface (SCI) is an asynchronous serial device available on the HC12. It can be either polled or interrupt driven and is intended for communication between remote devices. Related to SCI is the serial peripheral interface (SPI). SPI is a synchronous serial interface. It is intended for communication between units that support SPI like a network of multiple microcontrollers. Because of the synchronization of timing that is required, SPI uses a system of master/slave relationships between microcontrollers.

The pulse width modulation (PWM) subsystem is often used for motor and solenoid control. Using registers that are mapped to both the PWM unit and the microprocessor, a PWM output can be commanded by setting values for the period and duty cycle in the proper registers. This will result in a particular on-time and off-time voltage command.

Last, the serial in-circuit debugger (SDI) allows the microcontroller to connect to a PC for checking and modifying embedded software.

Microcontroller Network Systems

There is one last topic that should be mentioned in this section on inputs and outputs. Mechatronic systems often work with other systems in a network. Data and commands are thus transmitted from one system to another. While there are many different protocols, both open and proprietary, that could be mentioned about this networking, two will serve our purposes. The first is the manufacturing automation protocol (MAP) that was developed by General Motors Corporation. This system is based on the ISO Open Systems Interconnection (OSI) model and is especially designed for computer integrated manufacturing (CIM) and multiple PLCs. The second is the controller area network (CAN). This standard for serial communications was developed by Robert Bosch GmbH for use among embedded systems in a car.

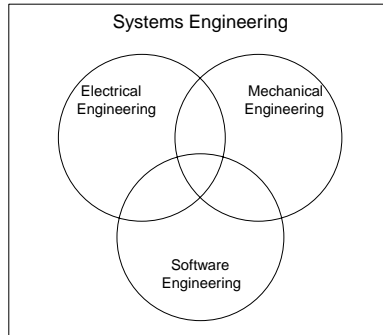



FIGURE 3.10 Mechatronics engineering disciplines.

3.8 Software Control

Systems Engineering

Systems engineering is the systems approach to the design and development of products and systems. As shown in Fig. 3.10, a drawing that shows the relationships of the major engineering competencies with mechatronics, the systems engineering competency encompasses the mechanical, electrical, and software competencies. There are several important tasks for the systems engineers to perform, starting with requirements gathering and continuing through final product and system verification and validation. After requirements gathering and analysis, the systems engineers should partition requirements functionality between mechanical, electrical, and software components, in consultation with the three competencies involved. This is part of the implementation of concurrent engineering. As also shown by the figure, software is an equal partner in the development of a mechatronic system. It is not an add-on to the system and it is not free, the two opinions that were sometimes held in the past by engineering management. While the phrase "Hardware adds cost, software adds value" is not entirely true either, sometimes software engineers felt that their competency was not given equal weight with the traditional engineering disciplines. And one last comment—many mechatronic systems are safety related, such as an air bag system in a car. It is as important for the software to be as fault tolerant as the hardware.

Software Engineering

Software engineering is concerned with both the final mechatronic "product" and the mechatronic development process. Two basic approaches are used with process, with many variations upon these approaches. One is called the "waterfall" method, where the process moves (falls) from one phase to another (e.g., analysis to design) with checkpoints along the way. The other method, the "spiral" approach, is often used when the requirements are not as well fixed. In this method there is prototyping, where the customers and/or systems engineers refine requirements as more information about the system becomes known. In either approach, once the requirements for the software portion of the mechatronic system are documented, the software engineers should further partition functionality as part of software design. Metrics as to development time, development cost, memory usage, and throughput should also be projected and recorded. Here is where the Software Engineering Institute's Capability Maturity Model (SEI CMM) levels can be used for guidance. It is a truism that software is almost never developed as easily as estimated, and that a system can remain at the "90% complete" level for most of the development life cycle. The first solution attempted to solve this problem is often assigning more software engineers onto the project. This does not always work, however, because of the learning curve of the new people, as stated by Frederick Brooks in his important book *The Mythical Man Month* (Addison-Wesley 1995).

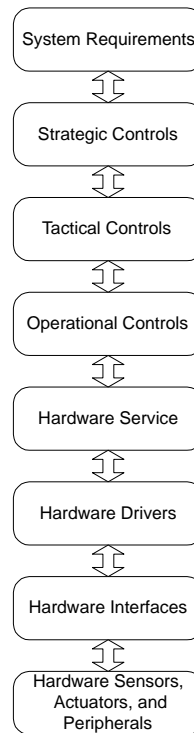



FIGURE 3.11 Mechatronic software layering.

Software Design

Perhaps the most important part of the software design for a mechatronic system can be seen from the hierarchy in Fig. 3.11. Ranging from requirements at the top to hardware at the bottom, this layering serves several purposes. The most important is that it separates mechatronic functionality from implementation. Quite simply, an upper layer should not be concerned with how a lower layer is actually performing a task. Each layer instead is directed by the layer above and receives a service or status from a layer below it. To cross more than one layer boundary is bad technique and can cause problems later in the process. Remember that this process abstraction is quite useful, for a mechatronic system has mechanical, electrical, and software parts all in concurrent development. A change in a sensor or actuator interface should only require a change at the layer immediately above, the driver layer. There is one last reason for using a hierarchical model such as this. In the current business climate, it is unlikely that the people working at the various layers will be collocated. Instead, it is not uncommon for development to be taking place in multiple locations in multiple countries. Without a crisp division of these layers, chaos can result.

For more information on these and many other topics in software engineering such as coupling, cohesion, and software reuse, please refer to Chapter 49 of this handbook, Roger Pressman's book *Software Engineering: A Practitioner's Approach 5th Edition* (McGraw Hill 2000), and Steve McConnell's book *Code Complete* (Microsoft Press 1993).

3.9 Testing and Instrumentation

Verification and Validation

Verification and validation are related tasks that should be completed throughout the life cycle of the mechatronic product or system. Boehm in his book *Software Engineering Economics* (Prentice-Hall 1988) describes verification as "building the product right" while validation is "building the right product." In other words, verification is the testing of the software and product to make sure that it is built to the design. Validation, on the other hand, is to make sure the software or product is built to the requirements

from the customer. As mentioned, verification and validation are life cycle tasks, not tasks completed just before the system is set for production. One of the simplest and most useful techniques is to hold hardware and software validation and verification reviews. Validation design reviews of hardware and software should include the systems engineers who have the best understanding of the customer requirements. Verification hardware design and software code reviews, or peer reviews, are an excellent means of finding errors upstream in the development process. Managers may have to decide whether to allocate resources upstream, when the errors are easier to fix, or downstream, when the ramifications can be much more drastic. Consider the difference between a code review finding a problem in code, and having the author change it and recompile, versus finding a problem after the product has been sold and in the field, where an expensive product recall may be required.

Debuggers

Edsgar Dijkstra, a pioneer in the development of programming as a discipline, discouraged the terms "bug" and "debug," and considered such terms harmful to the status of software engineering. They are, however, used commonly in the field. A debugger is a software program that allows a view of what is happening with the program code and data while the program is executing. Generally it runs on a PC that is connected to a special type of development microcontroller called an emulator. While debuggers can be quite useful in finding and correcting errors in code, they are not real-time, and so can actually create computer operating properly (COP) errors. However, if background debug mode (BDM) is available on the microprocessor, the debugger can be used to step through the algorithm of the program, making sure that the code is operating as expected. Intermediate and final variable values, especially those related to some analog input or output value, can be checked. Most debuggers allow multiple open windows, the setting of program execution break points in the code, and sometimes even the reflashing of the program into the microcontroller emulator. An example is the Noral debugger available for the Motorola HC12.

The software in the microcontroller can also check itself and its hardware. By programming in a checksum, or total, of designated portions of ROM and/or EEPROM, the software can check to make sure that program and data are correct. By alternately writing and reading 0x55 and 0xAA to RAM (the "checkerboard test"), the program can verify that RAM and the bus are operating properly. These startup tasks should be done with every product operation cycle.

Logic Analyzer

A logic analyzer is a device for nonintrusive monitoring and testing of the microcontroller. It is usually connected to both the microcontroller and a simulator. While the microcontroller is running its program and processing data, the simulator is simulating inputs and displaying outputs of the system. A "trigger word" can be entered into the logic analyzer. This is a bit pattern that will be on one of the buses monitored by the logic analyzer. With this trigger, the bus traffic around that point of interest can be captured and stored in the memory of the analyzer. An inverse assembler in the analyzer allows the machine code on the bus to be seen and analyzed in the form of the assembly level commands of the program. The analyzer can also capture the analog outputs of the microcontroller. This could be used to verify that the correct PWM duty cycle is being commanded. The simulator can introduce shorts or opens into the system, then the analyzer is used to see if the software correctly responds to the faults. The logic analyzer can also monitor the master loop of the system, making sure that the system completes all of its tasks within a designated time, e.g., 15 ms. An example of a logic analyzer is the Hewlett Packard HP54620.

3.10 Summary

This chapter introduced a number of topics regarding a mechatronic system. These topics included not just mechatronic input, output, and processing, but also design, development, and testing. Future chapters will cover all of this material in much greater detail.

4

Microprocessor-Based Controllers and Microelectronics

Ondrej Novak
Technical University Liberec

Ivan Dolezal
Technical University Liberec

4.1 Introduction to Microelectronics

The field of microelectronics has changed dramatically during the last two decades and digital technology has governed most of the application fields in electronics. The design of digital systems is supported by thousands of different integrated circuits supplied by many manufacturers across the world. This makes both the design and the production of electronic products much easier and cost effective. The permanent growth of integrated circuit speed, scale of integration, and reduction of costs have resulted in digital circuits being used instead of classical analog solutions of controllers, filters, and (de)modulators.

The growth in computational power can be demonstrated with the following example. One single-chip microcontroller has the computational power equal to that of one 1992 vintage computer notebook. This single-chip microcontroller has the computational power equal to four 1981 vintage IBM personal computers, or to two 1972 vintage IBM 370 mainframe computers.

Digital integrated circuits are designed to be universal and are produced in large numbers. Modern integrated circuits have many upgraded features from earlier designs, which allow for "user-friendlier" access and control. As the parameters of Integrated circuits (ICs) influence not only the individually designed IC, but all the circuits that must cooperate with it, a roadmap of the future development of IC technology is updated every year. From this roadmap we can estimate future parameters of the ICs, and adapt our designs to future demands. The relative growth of the number of integrated transistors on a chip is relatively stable. In the case of memory elements, it is equal to approximately 1.5 times the current amount. In the case of other digital ICs, it is equal to approximately 1.35 times the current amount.

In digital electronics, we use quantities called logical values instead of the analog quantities of voltage and current. Logical variables usually correspond to the voltage of the signal, but they have only two values: log.1 and log.0. If a digital circuit processes a logical variable, a correct value is recognized because between the logical value voltages there is a gap (see Fig. 4.1). We can arbitrarily improve the resolution of signals by simply using more bits.

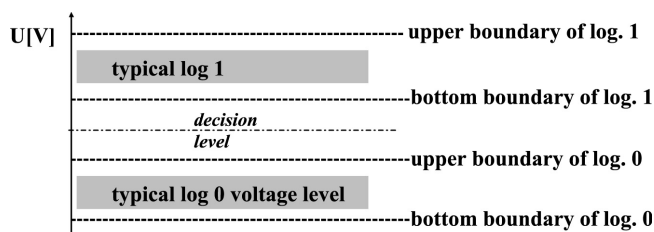



FIGURE 4.1 Voltage levels and logical values correspondence.

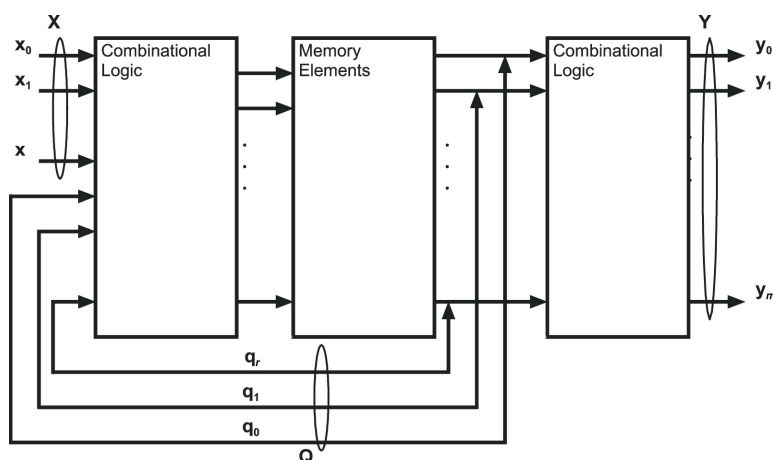



FIGURE 4.2 A finite state automaton: *X*—input binary vector, *Y*—output binary vector, *Q*—internal state vector.

4.2 Digital Logic

Digital circuits are composed of logic gates, such as elementary electronic circuits operating in only two states. These gates operate in such a way that the resulting logical value corresponds to the resulting value of the Boolean algebra statements. This means that with the help of gates we can realize every logical and arithmetical operation. These operations are performed in combinational circuits for which the resulting value is dependent only on the actual state of the inputs variables. Of course, logic gates are not enough for automata construction. For creating an automaton, we also need some memory elements in which we capture the responses of the arithmetical and logical blocks.

A typical scheme of a digital finite state automaton is given in Fig. 4.2. The automata can be constructed from standard ICs containing logic gates, more complex combinational logic blocks and registers, counters, memories, and other standard sequential ICs assembled on a printed circuit board. Another possibility is to use application specific integrated circuits (ASIC), either programmable or full custom, for a more advanced design. This approach is suitable for designs where fast hardware solutions are preferred. Another possibility is to use microcontrollers that are designed to serve as universal automata, which function can be specified by memory programming.

4.3 Overview of Control Computers

Huge, complex, and power-consuming single-room mainframe computers and, later, single-case minicomputers were primarily used for scientific and technical computing (e.g., in FORTRAN, ALGOL) and for database applications (e.g., in COBOL). The invention in 1971 of a universal central processing unit (CPU) in a single chip microprocessor caused a revolution in the computer technology. Beginning in

FIGURE 4.3 Example of a small mechatronic system: The ALAMBETA device for measurement of thermal properties of fabrics and plastic foils (manufactured by SENSORA, Czech Republic). It employs a unique measuring method using extra thin heat flow sensors, sample thickness measurement incorporated into a head drive, microprocessor control, and connection with a PC.

1981, multi-boxes (desktop or tower case, monitor, keyboard, mouse) or single-box (notebook) microcomputers became a daily-used personal tool for word processing, spreadsheet calculation, game playing, drawing, multimedia processing, and presentations. When connected in a local area network (LAN) or over the Internet, these "personal computers (PCs)" are able to exchange data and to browse the World Wide Web (WWW).

Besides these "visible" computers, many embedded microcomputers are hidden in products such as machines, vehicles, measuring instruments, telecommunication devices, home appliances, consumer electronic products (cameras, hi-fi systems, televisions, video recorders, mobile phones, music instruments, toys, air-conditioning). They are connected with sensors, user interfaces (buttons and displays), and actuators. Programmability of such controllers brings flexibility to the devices (function program choice), some kind of intelligence (fuzzy logic), and user-friendly action. It ensures higher reliability and easier maintenance, repairs, (auto)calibration, (auto)diagnostics, and introduces the possibility of their interconnection—mutual communication or hierarchical control in a whole plant or in a smart house. A photograph of an electrically operated instrument is given in Fig. 4.3.

Embedded microcomputers are based on the Harvard architecture where code and data memories are split. Firmware (program code) is cross-compiled on a development system and then resides in a nonvolatile memory. In this way, a single main program can run immediately after a supply is switched on. Relatively expensive and shock sensitive mechanical memory devices (hard disks) and vacuum tube monitors have been replaced with memory cards or solid state disks (if an archive memory is essential) and LED segment displays or LCDs. A PC-like keyboard can be replaced by a device/function specifically labeled key set and/or common keys (arrows, Enter, Escape) completed with numeric keys, if necessary. Such key sets, auxiliary switches, large buttons, the main switch, and display can be located in water and dust resistant operator panels.

Progress in circuit integration caused fast development of microcontrollers in the last two decades. Code memory, data memory, clock generator, and a diverse set of peripheral circuits are integrated with the CPU (Fig. 4.4) to insert such complete single-chip microcomputers into an application specific PCB.

Digital signal processors (DSPs) are specialized embedded microprocessors with some on-chip peripherals but with external ADC/DAC, which represent the most important input/output channel. DSPs have a parallel computing architecture and a fixed point or floating point instruction set optimized for typical signal processing operations such as discrete transformations, filtering, convolution, and coding. We can find DSPs in applications like sound processing/generation, sensor (e.g., vibration) signal analysis,

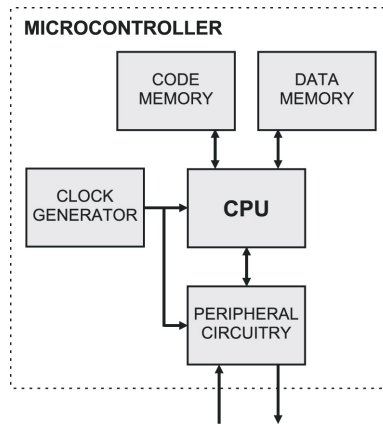



FIGURE 4.4 Block diagram of a microcontroller.

telecommunications (e.g., bandpass filter and digital modulation/demodulation in mobile phones, communication transceivers, modems), and vector control of AC motors.

Mass production (i.e., low cost), wide-spread knowledge of operation, comprehensive access to software development and debugging tools, and millions of ready-to-use code lines make PCs useful for computing-intensive measurement and control applications, although their architecture and operating systems are not well suited for this purpose.

As a result of computer expansion, there exists a broad spectrum of computing/processing means from powerful workstations, top-end PCs and VXI systems (64/32 bits, over 1000 MFLOPS/MIPS, 1000 MB of memory, input power over 100 W, cost about $10,000), downwards to PC-based computer cards/modules (32 bits, 100–300 MFLOPS/MIPS, 10–100 MB, cost less than $1000). Microprocessor cards/modules (16/8 bits, 10–30 MIPS, 1 MB, cost about $100), complex microcontroller chips (16/8 bits, 10–30 MIPS, 10–100 KB, cost about $10), and simple 8-pin microcontrollers (8 bits, 1–5 MIPS, 1 KB, 10 mW, cost about $1) are also available for very little money.

4.4 Microprocessors and Microcontrollers

There is no strict border between microprocessors and microcontrollers because certain chips can access external code and/or data memory (microprocessor mode) and are equipped with particular peripheral components.

Some microcontrollers have an internal RC oscillator and do not need an external component. However, an external quartz or ceramic resonator or RC network is frequently connected to the built-in, active element of the clock generator. Clock frequency varies from 32 kHz (extra low power) up to 75 MHz. Another auxiliary circuit generates the reset signal for an appropriate period after a supply is turned on. Watchdog circuits generate chip reset when a periodic retriggering signal does not come in time due to a program problem. There are several modes of consumption reduction activated by program instructions.

Complexity and structure of the interrupt system (total number of sources and their priority level selection), settings of level/edge sensitivity of external sources and events in internal (i.e., peripheral) sources, and handling of simultaneous interrupt events appear as some of the most important criteria of microcontroller taxonomy.

Although 16- and 32-bit microcontrollers are engaged in special, demanding applications (servo-unit control), most applications employ 8-bit chips. Some microcontrollers can internally operate with a 16-bit or even 32-bit data only in fixed-point range—microcontrollers are not provided with floating point unit (FPU). New microcontroller families are built on RISC (Reduced Instruction Set) core executing due to pipelining one instruction per few clock cycles or even per each cycle.

One can find further differences in addressing modes, number of direct accessible registers, and type of code memory (ranging from 1 to 128 KB) that are important from the view of firmware development. Flash memory enables quick and even in-system programming (ISP) using 3–5 wires, whereas classical EPROM makes chips more expensive due to windowed ceramic packaging. Some microcontrollers have built-in boot and debug capability to load code from a PC into the flash memory using UART (Universal Asynchronous Receiver/Transmitter) and RS-232C serial line. OTP (One Time Programmable) EPROM or ROM appear effective for large production series. Data EEPROM (from 64 B to 4 KB) for calibration constants, parameter tables, status storage, and passwords that can be written by firmware stand beside the standard SRAM (from 32 B to 4 KB).

The range of peripheral components is very wide. Every chip has bidirectional I/O (input/output) pins associated in 8-bit ports, but they often have an alternate function. Certain chips can set an input decision level (TTL, MOS, or Schmitt trigger) and pull-up or pull-down current sources. Output drivers vary in open collector or tri-state circuitry and maximal currents.

At least one 8-bit timer/counter (usually provided with a prescaler) counts either external events (optional pulses from an incremental position sensor) or internal clocks, to measure time intervals, and periodically generates an interrupt or variable baud rate for serial communication. General purpose 16-bit counters and appropriate registers form either capture units to store the time of input transients or compare units that generate output transients as a stepper motor drive status or PWM (pulse width modulation) signal. A real-time counter (RTC) represents a special kind of counter that runs even in sleep mode. One or two asynchronous and optionally synchronous serial interfaces (UART/USART) communicate with a master computer while other serial interfaces like SPI, CAN, and I^2C control other specific chips employed in the device or system.

Almost every microcontroller family has members that are provided with an A/D converter and a multiplexer of single-ended inputs. Input range is usually unipolar and equal to supply voltage or rarely to the on-chip voltage reference. The conversion time is given by the successive approximation principle of ADC, and the effective number of bits (ENOB) usually does not reach the nominal resolution 8, 10, or 12 bits.

There are other special interface circuits, such as field programmable gate array (FPGA), that can be configured as an arbitrary digital circuit.

Microcontroller firmware is usually programmed in an assembly language or in C language. Many software tools, including chip simulators, are available on websites of chip manufacturers or third-party companies free of charge. A professional integrated development environment and debugging hardware (in-circuit emulator) is more expensive (thousands of dollars). However, smart use of an inexpensive ROM simulator in a microprocessor system or a step-by-step development cycle using an ISP programmer of flash microcontroller can develop fairly complex applications.

4.5 Programmable Logic Controllers

A programmable logic controller (PLC) is a microprocessor-based control unit designed for an industrial installation (housing, terminals, ambient resistance, fault tolerance) in a power switchboard to control machinery or an industrial process. It consists of a CPU with memories and an I/O interface housed either in a compact box or in modules plugged in a frame and connected with proprietary buses. The compact box starts with about 16 I/O interfaces, while the module design can have thousands of I/O interfaces. Isolated inputs usually recognize industrial logic, 24 V DC or main AC voltage, while outputs are provided either with isolated solid state switches (24 V for solenoid valves and contactors) or with relays. Screw terminal boards represent connection facilities, which are preferred in PLCs to wire them to the controlled systems. I/O logical levels can be indicated with LEDs near to terminals.

Since PLCs are typically utilized to replace relays, they execute Boolean (bit, logical) operations and timer/counter functions (a finite state automaton). Analog I/O, integer or even floating point arithmetic, PWM outputs, and RTC are implemented in up-to-date PLCs. A PLC works by continually scanning a program, such as machine code, that is interpreted by an embedded microprocessor (CPU). The scan time is the time it takes to check the input status, to execute all branches (all individual rungs of a ladder

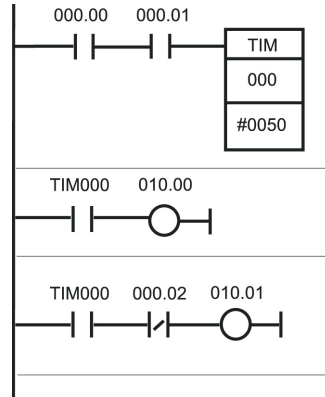



FIGURE 4.5 Example of PLC ladder diagram: 000.xx/010.xx—address group of inputs/outputs, TIM000—timer delays 5 s. 000.00—normally open input contact, 000.02—normally closed input contact.

diagram) of the program using internal (state) bit variables if any, and to update the output status. The scan time is dependent on the complexity of the program (milliseconds or tens of msec). The next scan operation either follows the previous one immediately (free running) or starts periodically.

Programming languages for PLCs are described in IEC-1131-3 nomenclature:

LD—ladder diagram (see Fig. 4.5)
IL—instruction list (an assembler)
SFC—sequential function chart (usually called by the proprietary name GRAFCET)
ST—structured text (similar to a high level language)
FBD—function block diagram

PLCs are programmed using cross-compiling and debugging tools running on a PC or with programming terminals (usually using IL), both connected with a serial link. Remote operator panels can serve as a human-to-machine interface. A new alternate concept (called SoftPLC) consists of PLC-like I/O modules controlled by an industrial PC, built in a touch screen operator panel.

4.6 Digital Communications

Intercommunication among mechatronics subsystems plays a key role in their engagement of applications, both of fixed and flexible configuration (a car, a hi-fi system, a fixed manufacturing line versus a flexible plant, a wireless pico-net of computer peripheral devices). It is clear that digital communication depends on the designers demands for the amount of transferred data, the distance between the systems, and the requirements on the degree of data reliability and security.

The signal is represented by alterations of amplitude, frequency, or phase. This is accomplished by changes in voltage/current in metallic wires or by electromagnetic waves, both in radiotransmission and infrared optical transmission (either "wireless" for short distances or optical fibers over fairly long distances). Data rate or bandwidth varies from 300 b/s (teleprinter), 3.4 kHz (phone), 144 kb/s (ISDN) to tens of Mb/s (ADSL) on a metallic wire (subscriber line), up to 100 Mb/s on a twisted pair (LAN), about 30–100 MHz on a microwave channel, 1 GHz on a coaxial cable (trunk cable network, cable TV), and up to tens of Gb/s on an optical cable (backbone network).

Data transmission employs complex methods of digital modulation, data compression, and data protection against loss due to noise interference, signal distortion, and dropouts. Multilayer standard protocols (ISO/OSI 7-layer reference model or Internet 4-layer group of protocols including well-known TCP/IP), "partly hardware, partly software realized," facilitate an understanding between communication systems. They not only establish connection on a utilizable speed, check data transfer, format and compress data, but can make communication transparent for an application. For example, no difference can be seen between local and remote data sources. An example of a multilayer communication concept is depicted in Fig. 4.6.

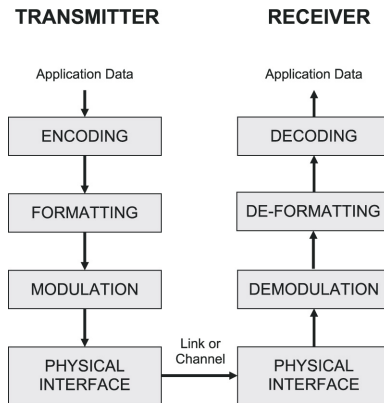



FIGURE 4.6 Example of multilayer communication.

Depending on the number of users, the communication is done either point-to-point (RS-232C from PC COM port to an instrument), point-to-multipoint (buses, networks), or even as a broadcasting (radio). Data are transferred using either switched connection (telephone network) or packet switching (computer networks, ATM). Bidirectional transmission can be full duplex (phone, RS-232C) or semi-duplex (most of digital networks). Concerning the link topology, a star connection or a tree connection employs a device ("master") mastering communication in the main node(s). A ring connection usually requires Token Passing method and a bus communication is controlled with various methods such as Master-Slave pooling, with or without Token Passing, or by using an indeterministic access (CSMA/CD in Ethernet).

An LPT PC port, SCSI for computer peripherals, and GPIB (IEEE-488) for instrumentation serve as examples of parallel (usually 8-bit) communication available for shorter distances (meters). RS-232C, RS-485, I^2C, SPI, USB, and Firewire (IEEE-1394) represent serial communication, some of which can bridge long distance (up to 1 km). Serial communication can be done either asynchronously using start and stop bits within transfer frame or synchronously using included synchronization bit patterns, if necessary. Both unipolar and bipolar voltage levels are used to drive either unbalanced lines (LPT, GPIB vs. RS-232C) or balanced twisted-pair lines (CAN vs. RS-422, RS-485).

5

An Introduction to Micro- and Nanotechnology

Michael Goldfarb
Vanderbilt University

Alvin Strauss
Vanderbilt University

Eric J. Barth
Vanderbilt University

5.1 Introduction

Originally arising from the development of processes for fabricating microelectronics, micro-scale devices are typically classified according not only to their dimensional scale, but their composition and manufacture. Nanotechnology is generally considered as ranging from the smallest of these micro-scale devices down to the assembly of individual molecules to form molecular devices. These two distinct yet overlapping fields of microelectromechanical systems (MEMS) and nanosystems or nanotechnology share a common set of engineering design considerations unique from other more typical engineering systems. Two major factors distinguish the existence, effectiveness, and development of micro-scale and nano-scale transducers from those of conventional scale. The first is the physics of scaling and the second is the suitability of manufacturing techniques and processes. The former is governed by the laws of physics and is thus a fundamental factor, while the latter is related to the development of manufacturing technology, which is a significant, though not fundamental, factor. Due to the combination of these factors, effective micro-scale transducers can often not be constructed as geometrically scaled-down versions of conventional-scale transducers.

The Physics of Scaling

The dominant forces that influence micro-scale devices are different from those that influence their conventional-scale counterparts. This is because the size of a physical system bears a significant influence on the physical phenomena that dictate the dynamic behavior of that system. For example, larger-scale systems are influenced by inertial effects to a much greater extent than smaller-scale systems, while smaller systems are influenced more by surface effects. As an example, consider small insects that can stand on the surface of still water, supported only by surface tension. The same surface tension is present when

humans come into contact with water, but on a human scale the associated forces are typically insignificant. The world in which humans live is governed by the same forces as the world in which these insects live, but the forces are present in very different proportions. This is due in general to the fact that inertial forces typically act in proportion to volume, and surface forces typically in proportion to surface area. Since volume varies with the third power of length and area with the second, geometrically similar but smaller objects have proportionally more area than larger objects.

Exact scaling relations for various types of forces can be obtained by incorporating dimensional analysis techniques [1–5]. Inertial forces, for example, can be dimensionally represented as $F_i = \rho L^3 \ddot{x}$, where F_i is a generalized inertia force, ρ is the density of an object, L is a generalized length, and x is a displacement. This relationship forms a single dimensionless group, given by

$$\Pi = \frac{F_i}{\rho L^3 \ddot{x}}$$

Scaling with geometric and kinematic similarity can be expressed as

$$\frac{L_s}{L_o} = \frac{x_s}{x_o} = N, \qquad \frac{t_s}{t_o} = 1$$

where L represents the length scale, x the kinematic scale, t the time scale, the subscript o the original system, and the s represents the scaled system. Since physical similarity requires that the dimensionless group (Π) remain invariant between scales, the force relationship is given by $F_s/F_o = N^4$, assuming that the intensive property (density) remains invariant (i.e., $\rho_s = \rho_o$). An inertial force thus scales as N^4, where N is the geometric scaling factor. Alternately stated, for an inertial system that is geometrically smaller by a factor of N, the force required to produce an equivalent acceleration is smaller by a factor of N^4. A similar analysis shows that viscous forces, dimensionally represented by $F_v = \mu L \dot{x}$, scale as N^2, assuming the viscosity μ remains invariant, and elastic forces, dimensionally represented by $F_e = ELx$, scale as N^2, assuming the elastic modulus E remains invariant. Thus, for a geometrically similar but smaller system, inertial forces will become considerably less significant with respect to viscous and elastic forces.

General Mechanisms of Electromechanical Transduction

The fundamental mechanism for both sensing and actuation is energy transduction. The primary forms of physical electromechanical transduction can be grouped into two categories. The first is multicomponent transduction, which utilizes "action at a distance" behavior between multiple bodies, and the second is deformation-based or solid-state transduction, which utilizes mechanics-of-material phenomena such as crystalline phase changes or molecular dipole alignment. The former category includes electromagnetic transduction, which is typically based upon the Lorentz equation and Faraday's law, and electrostatic interaction, which is typically based upon Coulomb's law. The latter category includes piezoelectric effects, shape memory alloys, and magnetostrictive, electrostrictive, and photostrictive materials. Although materials exhibiting these properties are beginning to be seen in a limited number of research applications, the development of micro-scale systems is currently dominated by the exploitation of electrostatic and electromagnetic interactions. Due to their importance, electrostatic and electromagnetic transduction is treated separately in the sections that follow.

Sensor and Actuator Transduction Characteristics

Characteristics of concern for both microactuator and microsensor technology are repeatability, the ability to fabricate at a small scale, immunity to extraneous influences, sufficient bandwidth, and if possible, linearity. Characteristics typically of concern specifically for microactuators are achievable force, displacement, power, bandwidth (or speed of response), and efficiency. Characteristics typically of concern specifically for microsensors are high resolution and the absence of drift and hysteresis.

5.2 Microactuators

Electrostatic Actuation

The most widely utilized multicomponent microactuators are those based upon electrostatic transduction. These actuators can also be regarded as a variable capacitance type, since they operate in an analogous mode to variable reluctance type electromagnetic actuators (e.g., variable reluctance stepper motors). Electrostatic actuators have been developed in both linear and rotary forms. The two most common configurations of the linear type of electrostatic actuators are the normal-drive and tangential or comb-drive types, which are illustrated in Figs. 5.1 and 5.2, respectively. Note that both actuators are suspended by flexures, and thus the output force is equal to the electrostatic actuation force minus the elastic force required to deflect the flexure suspension. The normal-drive type of electrostatic microactuator operates in a similar fashion to a condenser microphone. In this type of drive configuration, the actuation force is given by

$$F_x = \frac{\varepsilon A v^2}{2x^2}$$

where A is the total area of the parallel plates, ε is the permittivity of air, v is the voltage across the plates, and x is the plate separation. The actuation force of the comb-drive configuration is given by

$$F_x = \frac{\varepsilon w v^2}{2d}$$

where w is the width of the plates, ε is the permittivity of air, v is the voltage across the plates, and d is the plate separation. Dimensional examination of both relations indicates that force is independent of geometric and kinematic scaling, that is, for an electrostatic actuator that is geometrically and kinematically reduced by a factor of N, the force produced by that actuator will be the same. Since forces associated with most other physical phenomena are significantly reduced at small scales, micro-scale electrostatic forces become significant relative to other forces. Such an observation is clearly demonstrated by the fact that all intermolecular forces are electrostatic in origin, and thus the strength of all materials is a result of electrostatic forces [6].

The maximum achievable force of multicomponent electrostatic actuators is limited by the dielectric breakdown of air, which occurs in dry air at about 0.8×10^6 V/m. Fearing [7] estimates that the upper limit for force generation in electrostatic actuation is approximately 10 N/cm^2. Since electrostatic drives

FIGURE 5.1 Schematic of a normal-drive electrostatic actuator.

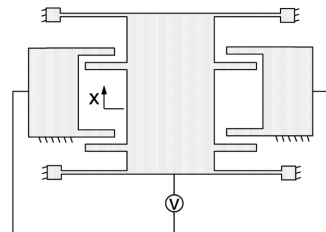
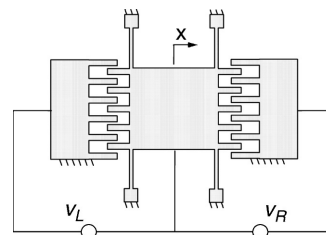



FIGURE 5.2 Comb-drive electrostatic actuator. Energizing an electrode provides motion toward that electrode.

do not have any significant actuation dynamics, and since the inertia of the moving member is usually small, the actuator bandwidth is typically quite large, on the order of a kilohertz.

The maximum achievable stroke for normal configuration actuators is limited by the elastic region of the flexure suspension and additionally by the dependence of actuation force on plate separation, as given by the above stated equations. According to Fearing, a typical stroke for a surface micromachined normal configuration actuator is on the order of a couple of microns. The achievable displacement can be increased by forming a stack of normal-configuration electrostatic actuators in series, as proposed by Bobbio et al. [8,9].

The typical stroke of a surface micromachined comb actuator is on the order of a few microns, though sometimes less. The maximum achievable stroke in a comb drive is limited primarily by the mechanics of the flexure suspension. The suspension should be compliant along the direction of actuation to enable increased displacement, but must be stiff orthogonal to this direction to avoid parallel plate contact due to misalignment. These modes of behavior are unfortunately coupled, so that increased compliance along the direction of motion entails a corresponding increase in the orthogonal direction. The net effect is that increased displacement requires increased plate separation, which results in decreased overall force.

The most common configurations of rotary electrostatic actuators are the variable capacitance motor and the wobble or harmonic drive motor, which are illustrated in Figs. 5.3 and 5.4, respectively. Both motors operate in a similar manner to the comb-drive linear actuator. The variable capacitance motor is characterized by high-speed low-torque operation. Useful levels of torque for most applications therefore require some form of significant micromechanical transmission, which do not presently exist. The rotor of the wobble motor operates by rolling along the stator, which provides an inherent harmonic-drive-type transmission and thus a significant transmission ratio (on the order of several hundred times). Note that the rotor must be well insulated to roll along the stator without electrical contact. The drawback to this approach is that the rotor motion is not concentric with respect to the stator, which makes the already difficult problem of coupling a load to a micro-shaft even more difficult.

Examples of normal type linear electrostatic actuators are those by Bobbio et al. [8,9] and Yamaguchi et al. [10]. Examples of comb-drive electrostatic actuators are those by Kim et al. [11] and Matsubara et al. [12], and a larger-scale variation by Niino et al. [13]. Examples of variable capacitance rotary electrostatic motors are those by Huang et al. [14], Mehragany et al. [15], and Trimmer and Gabriel [16].

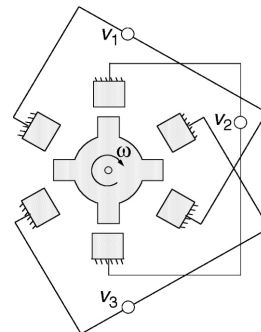



FIGURE 5.3 Variable capacitance type electrostatic motor. Opposing pairs of electrodes are energized sequentially to rotate the rotor.

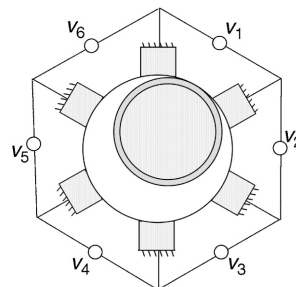



FIGURE 5.4 Harmonic drive type electrostatic motor. Adjacent electrodes are energized sequentially to roll the (insulated) rotor around the stator.

Examples of harmonic-drive motors are those by Mehragany et al. [17,18], Price et al. [19], Trimmer and Jebens [20,21], and Furuhata et al. [22]. Electrostatic microactuators remain a subject of research interest and development, and as such are not yet available on the general commercial market.

Electromagnetic Actuation

Electromagnetic actuation is not as omnipresent at the micro-scale as at the conventional-scale. This probably is due in part to early skepticism regarding the scaling of magnetic forces, and in part to the fabrication difficulty in replicating conventional-scale designs. Most electromagnetic transduction is based upon a current carrying conductor in a magnetic field, which is described by the Lorentz equation:

$$dF = Idl \times B$$

where F is the force on the conductor, I is the current in the conductor, l is the length of the conductor, and B is the magnetic flux density. In this relation, the magnetic flux density is an intensive variable and thus (for a given material) does not change with scale. Scaling of current, however, is not as simple. The resistance of wire is given by

$$R = \frac{\rho l}{A}$$

where ρ is the resistivity of the wire (an intensive variable), l is the length, and A the cross-sectional area. If a wire is geometrically decreased in size by a factor of N, its resistance will increase by a factor of N. Since the power dissipated in the wire is I^2R, assuming the current remains constant implies that the power dissipated in the geometrically smaller wire will increase by a factor of N. Assuming the maximum power dissipation for a given wire is determined by the surface area of the wire, a wire that is smaller by a factor of N will be able to dissipate a factor of N^2 less power. Constant current is therefore a poor assumption. A better assumption is that maximum current is limited by maximum power dissipation, which is assumed to depend upon surface area of the wire. Since a wire smaller by a factor of N can dissipate a factor of N^2 less power, the current in the smaller conductor would have to be reduced by a factor of $N^{3/2}$. Incorporating this into the scaling of the Lorentz equation, an electromagnetic actuator that is geometrically smaller by a factor of N would exert a force that is smaller by a factor of $N^{5/2}$. Trimmer and Jebens have conducted a similar analysis, and demonstrated that electromagnetic forces scale as N^2 when assuming constant temperature rise in the wire, $N^{5/2}$ when assuming constant heat (power) flow (as previously described), and N^3 when assuming constant current density [23,24]. In any of these cases, the scaling of electromagnetic forces is not nearly as favorable as the scaling of electrostatic forces. Despite this, electromagnetic actuation still offers utility in microactuation, and most likely scales more favorably than does inertial or gravitational forces.

Lorentz-type approaches to microactuation utilize surface micromachined micro-coils, such as the one illustrated in Fig. 5.5. One configuration of this approach is represented by the actuator of Inoue et al. [25],

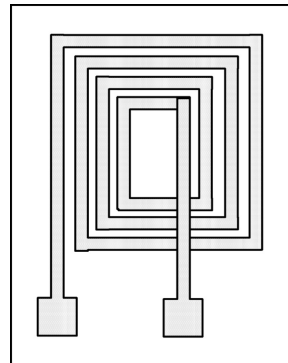



FIGURE 5.5 Schematic of surface micromachined microcoil for electromagnetic actuation.

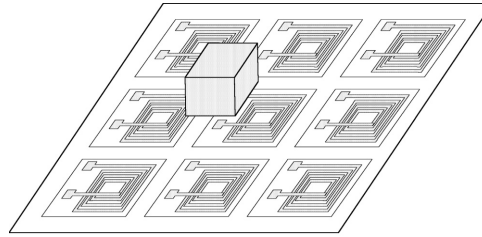



FIGURE 5.6 Microcoil array for planar positioning of a permanent micromagnet, as described by Inoue et al. [25]. Each coil produces a field, which can either attract or repel the permanent magnet, as determined by the direction of current. The magnet does not levitate, but rather slides on the insulated surface.

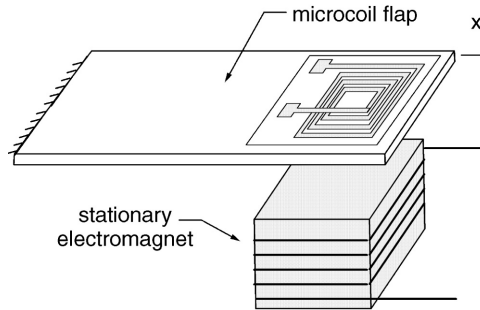



FIGURE 5.7 Cantilevered microcoil flap as described by Liu et al. [26]. The interaction between the energized coil and the stationary electromagnet deflects the flap upward or downward, depending on the direction of current through the microcoil.

which utilizes current control in an array of microcoils to position a permanent micro-magnet in a plane, as illustrated in Fig. 5.6. Another Lorentz-type approach is illustrated by the actuator of Liu et al. [26], which utilizes current control of a cantilevered microcoil flap in a fixed external magnetic field to effect deflection of the flap, as shown in Fig. 5.7. Liu reported deflections up to 500 μm and a bandwidth of approximately 1000 Hz [26]. Other examples of Lorentz-type nonrotary actuators are those by Shinozawa et al. [27], Wagner and Benecke [28], and Yanagisawa et al. [29]. A purely magnetic approach (i.e., not fundamentally electromagnetic) is the work of Judy et al. [30], which in essence manipulates a flexure-suspended permanent micromagnet by controlling an external magnetic field.

Ahn et al. [31] and Guckel et al. [32] have both demonstrated planar rotary variable-reluctance type electromagnetic micromotors. A variable reluctance approach is advantageous because the rotor does not require commutation and need not be magnetic. The motor of Ahn et al. incorporates a 12-pole stator and 10-pole rotor, while the motor of Guckel et al. utilizes a 6-pole stator and 4-pole rotor. Both incorporate rotors of approximately 500 μm diameter. Guckel reports (no load) rotor speeds above 30,000 rev/min, and Ahn estimates maximum stall torque at 1.2 μN m. As with electrostatic microactuators, microfabricated electromagnetic actuators likewise remain a subject of research interest and development and as such are not yet available on the general commercial market.

5.3 Microsensors

Since microsensors do not transmit power, the scaling of force is not typically significant. As with conventional-scale sensing, the qualities of interest are high resolution, absence of drift and hysteresis, achieving a sufficient bandwidth, and immunity to extraneous effects not being measured.

Microsensors are typically based on either measurement of mechanical strain, measurement of mechanical displacement, or on frequency measurement of a structural resonance. The former two types

are in essence analog measurements, while the latter is in essence a binary-type measurement, since the sensed quantity is typically the frequency of vibration. Since the resonant-type sensors measure frequency instead of amplitude, they are generally less susceptible to noise and thus typically provide a higher resolution measurement. According to Guckel et al., resonant sensors provide as much as one hundred times the resolution of analog sensors [33]. They are also, however, more complex and are typically more difficult to fabricate.

The primary form of strain-based measurement is piezoresistive, while the primary means of displacement measurement is capacitive. The resonant sensors require both a means of structural excitation as well as a means of resonant frequency detection. Many combinations of transduction are utilized for these purposes, including electrostatic excitation, capacitive detection, magnetic excitation and detection, thermal excitation, and optical detection.

Strain

Many microsensors are based upon strain measurement. The primary means of measuring strain is via piezoresistive strain gages, which is an analog form of measurement. Piezoresistive strain gages, also known as semiconductor gages, change resistance in response to a mechanical strain. Note that piezoelectric materials can also be utilized to measure strain. Recall that mechanical strain will induce an electrical charge in a piezoelectric ceramic. The primary problem with using a piezoelectric material, however, is that since measurement circuitry has limited impedance, the charge generated from a mechanical strain will gradually leak through the measurement impedance. A piezoelectric material therefore cannot provide reliable steady-state signal measurement. In constrast, the change in resistance of a piezoresistive material is stable and easily measurable for steady-state signals. One problem with piezoresistive materials, however, is that they exhibit a strong strain-temperature dependence, and so must typically be thermally compensated.

An interesting variation on the silicon piezoresistor is the resonant strain gage proposed by Ikeda et al., which provides a frequency-based form of measurement that is less susceptible to noise [34]. The resonant strain gage is a beam that is suspended slightly above the strain member and attached to it at both ends. The strain gage beam is magnetically excited with pulses, and the frequency of vibration is detected by a magnetic detection circuit. As the beam is stretched by mechanical strain, the frequency of vibration increases. These sensors provide higher resolution than typical piezoresistors and have a lower temperature coefficient. The resonant sensors, however, require a complex three-dimensional fabrication technique, unlike the typical piezoresistors which require only planar techniques.

Pressure

One of the most commercially successful microsensor technologies is the pressure sensor. Silicon micromachined pressure sensors are available that measure pressure ranges from around one to several thousand kPa, with resolutions as fine as one part in ten thousand. These sensors incorporate a silicon micromachined diaphragm that is subjected to fluid (i.e., liquid or gas) pressure, which causes dilation of the diaphragm. The simplest of these utilize piezoresistors mounted on the back of the diaphragm to measure deformation, which is a function of the pressure. Examples of these devices are those by Fujii et al. [35] and Mallon et al. [36]. A variation of this configuration is the device by Ikeda et al. Instead of a piezoresistor to measure strain, an electromagnetically driven and sensed resonant strain gage, as discussed in the previous section, is utilized [37]. Still another variation on the same theme is the capacitive measurement approach, which measures the capacitance between the diaphragm and an electrode that is rigidly mounted and parallel to the diaphragm. An example of this approach is by Nagata et al. [38]. A more complex approach to pressure measurement is that by Stemme and Stemme, which utilizes resonance of the diaphragm to detect pressure [39]. In this device, the diaphragm is capacitively excited and optically detected. The pressure imposes a mechanical load on the diaphragm, which increases the stiffness and, in turn, the resonant frequency.

Acceleration

Another commercially successful microsensor is the silicon microfabricated accelerometer, which in various forms can measure acceleration ranges from well below one to around a thousand meters per square second (i.e., sub-*g* to several hundred *g*'s), with resolutions of one part in 10,000. These sensors incorporate a micromachined suspended proof mass that is subjected to an inertial force in response to an acceleration, which causes deflection of the supporting flexures. One means of measuring the deflection is by utilizing piezoresistive strain gages mounted on the flexures. The primary disadvantage to this approach is the temperature sensitivity of the piezoresistive gages. An alternative to measuring the deflection of the proof mass is via capacitive sensing. In these devices, the capacitance is measured between the proof mass and an electrode that is rigidly mounted and parallel. Examples of this approach are those by Boxenhorn and Greiff [40], Leuthold and Rudolf [41], and Seidel et al. [42]. Still another means of measuring the inertial force on the proof mass is by measuring the resonant frequency of the supporting flexures. The inertial force due to acceleration will load the flexure, which will alter its resonant frequency. The frequency of vibration is therefore a measure of the acceleration. These types of devices utilize some form of transduction to excite the structural resonance of the supporting flexures, and then utilize some other measurement technique to detect the frequency of vibration. Examples of this type of device are those by Chang et al. [43], which utilize electrostatic excitation and capacitive detection, and by Satchell and Greenwood [44], which utilize thermal excitation and piezoresistive detection. These types of accelerometers entail additional complexity, but typically offer improved measurement resolution. Still another variation of the micro-accelerometer is the force-balanced type. This type of device measures position of the proof mass (typically by capacitive means) and utilizes a feedback loop and electrostatic or electromagnetic actuation to maintain zero deflection of the mass. The acceleration is then a function of the actuation effort. These devices are characterized by a wide bandwidth and high sensitivity, but are typically more complex and more expensive than other types. Examples of force-balanced devices are those by Chau et al. [45], and Kuehnel and Sherman [46], both of which utilize capacitive sensing and electrostatic actuation.

Force

Silicon microfabricated force sensors incorporate measurement approaches much like the microfabricated pressure sensors and accelerometers. Various forms of these force sensors can measure forces ranging on the order of millinewtons to newtons, with resolutions of one part in 10,000. Mechanical sensing typically utilizes a beam or a flexure support which is elastically deflected by an applied force, thereby transforming force measurement into measurement of strain or displacement, which can be accomplished by piezoresistive or capacitive means. An example of this type of device is that of Despont et al., which utilizes capacitive measurement [47]. Higher resolution devices are typically of the resonating beam type, in which the applied force loads a resonating beam in tension. Increasing the applied tensile load results in an increase in resonant frequency. An example of this type of device is that of Blom et al. [48].

Angular Rate Sensing (Gyroscopes)

A conventional-scale gyroscope utilizes the spatial coupling of the angular momentum-based gyroscopic effect to measure angular rate. In these devices, a disk is spun at a constant high rate about its primary axis, so that when the disk is rotated about an axis not colinear with the primary (or spin) axis, a torque results in an orthogonal direction that is proportional to the angular velocity. These devices are typically mounted in gimbals with low-friction bearings, incorporate motors that maintain the spin velocity, and utilize strain gages to measure the gyroscopic torque (and thus angular velocity). Such a design would not be appropriate for a microsensor due to several factors, some of which include the diminishing effect of inertia (and thus momentum) at small scales, the lack of adequate bearings, the lack of appropriate micromotors, and the lack of an adequate three-dimensional microfabrication processes. Instead, microscale angular rate sensors are of the vibratory type, which incorporate Coriolis-type effects rather than

FIGURE 5.8 Illustration of Coriolis acceleration, which results from translation within a reference frame that is rotating with respect to an inertial reference frame.

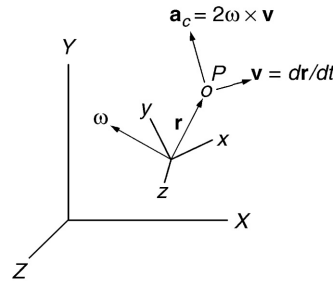
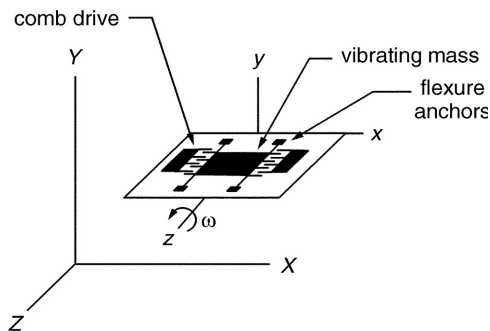



FIGURE 5.9 Schematic of a vibratory gyroscope.

the angular momentum-based gyroscopic mechanics of conventional-scale devices. A Coriolis acceleration results from linear translation within a coordinate frame that is rotating with respect to an inertial reference frame. In particular, if the particle in Fig. 5.8 is moving with a velocity v within the frame xyz, and if the frame xyz is rotating with an angular velocity of ω with respect to the inertial reference frame XYZ, then a Coriolis acceleration will result equal to $\mathbf{a}_c = 2\omega \times v$. If the object has a mass m, a Coriolis inertial force will result equal to $\mathbf{F}_c = -2m\omega \times v$ (minus sign because direction is opposite $\mathbf{a}_c$). A vibratory gyroscope utilizes this effect as illustrated in Fig. 5.9. A flexure-suspended inertial mass is vibrated in the x-direction, typically with an electrostatic comb drive. An angular velocity about the z-axis will generate a Coriolis acceleration, and thus force, in the y-direction. If the "external" angular velocity is constant and the velocity in the x-direction is sinusoidal, then the resulting Coriolis force will be sinusiodal, and the suspended inertial mass will vibrate in the y-direction with an amplitude proportional to the angular velocity. The motion in the y-direction, which is typically measured capacitively, is thus a measure of the angular rate. Examples of these types of devices are those by Bernstein et al. [49] and Oh et al. [50]. Note that though vibration is an essential component of these devices, they are not technically resonant sensors, since they measure amplitude of vibration rather than frequency.

5.4 Nanomachines

Nanomachines are devices that range in size from the smallest of MEMS devices down to devices assembled from individual molecules [51]. This section briefly introduces energy sources, structural hierarchy, and the projected future of the assembly of nanomachines. Built from molecular components performing individual mechanical functions, the candidates for energy sources to actuate nanomachines are limited to those that act on a molecular scale. Regarding manufacture, the assembly of nanomachines is by nature a one-molecule-at-a-time operation. Although microscopy techniques are currently used for the assembly of nanostructures, self-assembly is seen as a viable means of mass production.

In a molecular device a discrete number of molecular components are combined into a supramolecular structure where each discrete molecular component performs a single function. The combined action of these individual molecules causes the device to operate and perform its various functions. Molecular devices require an energy source to operate. This energy must ultimately be used to activate the component molecules in the device, and so the energy must be chemical in nature. The chemical energy can be obtained by adding hydrogen ions, oxidants, etc., by inducing chemical reactions by the impingement of light, or by the actions of electrical current. The latter two means of energy activation, photochemical and electrochemical energy sources, are preferred since they not only provide energy for the operation of the device, but they can also be used to locate and control the device. Additionally, such energy transduction can be used to transmit data to report on the performance and status of the device. Another reason for the preference for photochemical- and electrochemical-based molecular devices is that, as these devices are required to operate in a cyclic manner, the chemical reactions that drive the system must be reversible. Since photochemical and electrochemical processes do not lead to the accumulation of products of reaction, they readily lend themselves to application in nanodevices.

Molecular devices have recently been designed that are capable of motion and control by photochemical methods. One device is a molecular plug and socket system, and another is a piston-cylinder system [51]. The construction of such supramolecular devices belongs to the realm of the chemist who is adept at manipulating molecules.

As one proceeds upwards in size to the next level of nanomachines, one arrives at devices assembled from (or with) single-walled carbon nanotubes (SWNTs) and/or multi-walled carbon nanotubes (MWNTs) that are a few nanometers in diameter. We will restrict our discussion to carbon nanotubes (CNTs) even though there is an expanding database on nanotubes made from other materials, especially bismuth. The strength and versatility of CNTs make them superior tools for the nanomachine design engineer. They have high electrical conductivity with current carrying capacity of a billion amperes per square centimeter. They are excellent field emitters at low operating voltages. Moreover, CNTs emit light coherently and this provides for an entire new area of holographic applications. The elastic modulus of CNTs is the highest of all materials known today [52]. These electrical properties and extremely high mechanical strength make MWNTs the ultimate atomic force microscope probe tips. CNTs have the potential to be used as efficient molecular assembly devices for manufacturing nanomachines one atom at a time.

Two obvious nanotechnological applications of CNTs are nanobearings and nanosprings. Zettl and Cumings [53] have created MWNT-based linear bearings and constant force nanosprings. CNTs may potentially form the ultimate set of nanometer-sized building blocks, out of which nanomachines of all kinds can be built. These nanomachines can be used in the assembly of nanomachines, which can then be used to construct machines of all types and sizes. These machines can be competitive with, or perhaps surpass existing devices of all kinds.

SWNTs can also be used as electromechanical actuators. Baughman et al. [54] have demonstrated that sheets of SWNTs generate larger forces than natural muscle and larger strains than high-modulus ferroelectrics. They have predicted that actuators using optimized SWNT sheets may provide substantially higher work densities per cycle than any other known actuator. Kim and Lieber [55] have built SWNT and MWNT nanotweezers. These nanoscale electromechanical devices were used to manipulate and interrogate nanostructures. Electrically conducting CNTs were attached to electrodes on pulled glass micropipettes. Voltages applied to the electrodes opened and closed the free ends of the CNTs. Kim and Lieber demonstrated the capability of the nanotweezers by grabbing and manipulating submicron clusters and nanowires. This device could be used to manipulate biological cells or even manipulate organelles and clusters within human cells. Perhaps, more importantly, these tweezers can potentially be used to assemble other nanomachines.

A wide variety of nanoscale manipulators have been proposed [56] including pneumatic manipulators that can be configured to make tentacle, snake, or multi-chambered devices. Drexler has proposed telescoping nanomanipulators for precision molecular positioning and assembly work. His manipulator has a cylindrical shape with a diameter of 35 nm and an extensible length of 100 nm. A number of six

degree of freedom Stewart platforms have been proposed [56], including one that allows strut lengths to be moved in 0.10 nm increments across a 100 nm work envelope. A number of other nanodevices including box-spring accelerometers, displacement accelerometers, pivoted gyroscopic accelerometers, and gimbaled nanogyroscopes have been proposed and designed [56].

Currently, much thought is being devoted to molecular assembly and self-replicating devices (self-replicating nanorobots). Self-assembly is arguably the only way for nanotechnology to advance in an engineering or technological sense. Assembling a billion or trillion atom device—one atom at a time—would be a great accomplishment. It would take a huge investment in equipment, labor, and time. Freitas [56] describes the infrastructure needed to construct a simple medical nanorobot: a 1-μm spherical respirocyte consisting of about 18 billion atoms. He estimates that a factory production line deploying a coordinated system of 100 macroscale scanning probe microscope (SPM) assemblers, where each assembler is capable of depositing one atom per second on a convergently-assembled workpiece, would result in a manufacturing throughput of two nanorobots per decade. If one conjectures about enormous increases in assembler manufacturing rates even to the extent of an output of one nanorobot per minute, it would take two million years to build the first cubic centimeter therapeutic dosage of nanorobots. Thus, it is clear that the future of medical nanotechnology and nanoengineering lies in the direction of self-assembly and self-replication.

References

1. Bridgman, P. W., *Dimensional Analysis,* 2nd Ed., Yale University Press, 1931.
2. Buckingham, E., "On physically similar systems: illustrations of the use of dimensional equations," *Physical Review,* 4(4):345–376, 1914.
3. Huntley, H. E., *Dimensional Analysis,* Dover Publications, 1967.
4. Langhaar, H. L., *Dimensional Analysis and Theory of Models,* John Wiley and Sons, 1951.
5. Taylor, E. S., *Dimensional Analysis for Engineers,* Oxford University Press, 1974.
6. Israelachvili, J. N., *Intermolecular and Surface Forces,* Academic Press, 1985, pp. 9–10.
7. Fearing, R. S., "Microactuators for microrobots: electric and magnetic," *Workshop on Micromechatronics, IEEE International Conference on Robotics and Automation,* 1997.
8. Bobbio, S. M., Keelam, M. D., Dudley, B. W., Goodwin-Hohansson, S., Jones, S. K., Jacobson, J. D., Tranjan, F. M., Dubois, T. D., "Integrated force arrays," *Proceedings of the IEEE Micro Electro Mechanical Systems,* 149–154, 1993.
9. Jacobson, J. D., Goodwin-Johansson, S. H., Bobbio, S. M., Bartlett, C. A., Yadon, L. N., "Integrated force arrays: theory and modeling of static operation," *Journal of Microelectromechanical Systems,* 4(3):139–150, 1995.
10. Yamaguchi, M., Kawamura, S., Minami, K., Esashi, M., "Distributed electrostatic micro actuators," *Proceedings of the IEEE Micro Electro Mechanical Systems,* 18–23, 1993.
11. Kim, C. J., Pisano, A. P., Muller, R. S., "Silicon-processed overhanging microgripper," *Journal of Microelectromechanical Systems,* 1(1):31–36, 1992.
12. Matsubara, T., Yamaguchi, M., Minami, K., Esashi, M., "Stepping electrostatic microactuator," *International Conference on Solid-State Sensor and Actuators,* 50–53, 1991.
13. Niino, T., Egawa, S., Kimura, H., Higuchi, T., "Electrostatic artificial muscle: compact, high-power linear actuators with multiple-layer structures," *Proceedings of the IEEE Conference on Micro Electro Mechanical Systems,* 130–135, 1994.
14. Huang, J. B., Mao, P. S., Tong, Q. Y., Zhang, R. Q., "Study on silicon electrostatic and electroquasistatic micromotors," *Sensors and Actuators,* 35:171–174, 1993.
15. Mehragany, M., Bart, S. F., Tavrow, L. S., Lang, J. H., Senturia, S. D., Schlecht, M. F., "A study of three microfabricated variable-capacitance motors," *Sensors and Actuators,* 173–179, 1990.
16. Trimmer, W., Gabriel, K., "Design considerations for a practical electrostatic micromotor," *Sensors and Actuators,* 11:189–206, 1987.

17. Mehregany, M., Nagarkar, P., Senturia, S. D., Lang, J. H., "Operation of microfabricated harmonic and ordinary side-drive motors," *Proceeding of the IEEE Conference on Micro Electro Mechanical Systems,* 1–8, 1990.
18. Dhuler, V. R., Mehregany, M., Phillips, S. M., "A comparative study of bearing designs and operational environments for harmonic side-drive micromotors," *IEEE Transactions on Electron Devices,* 40(11):1985–1989, 1993.
19. Price, R. H., Wood, J. E., Jacobsen, S. C., "Modeling considerations for electrostatic forces in electrostatic microactuators," *Sensors and Actuators,* 20:107–114, 1989.
20. Trimmer, W., Jebens, R., "An operational harmonic electrostatic motor," *Proceeding of the IEEE Conference on Micro Electro Mechanical Systems,* 13–16, 1989.
21. Trimmer, W., Jebens, R., "Harmonic electrostatic motors," *Sensors and Actuators,* 20:17–24, 1989.
22. Furuhata, T., Hirano, T., Lane, L. H., Fontanta, R. E., Fan, L. S., Fujita, H., "Outer rotor surface micromachined wobble micromotor," *Proceeding of the IEEE Conference on Micro Electro Mechanical Systems,* 161–166, 1993.
23. Trimmer, W., Jebens, R., "Actuators for microrobots," *IEEE Conference on Robotics and Automation,* 1547–1552, 1989.
24. Trimmer, W., "Microrobots and micromechanical systems," *Sensors and Actuators,* 19:267–287, 1989.
25. Inoue, T., Hamasaki, Y., Shimoyama, I., Miura, H., "Micromanipulation using a microcoil array," *Proceedings of the IEEE International Conference on Robotics and Automation,* 2208–2213, 1996.
26. Liu, C., Tsao, T., Tai, Y., Ho, C., "Surface micromachined magnetic actuators," *Proceedings of the IEEE Conference on Micro Electro Mechanical Systems,* 57–62, 1994.
27. Shinozawa, Y., Abe, T., Kondo, T., "A proportional microvalve using a bi-stable magnetic actuator," *Proceedings of the IEEE Conference on Micro Electro Mechanical Systems,* 233–237, 1997.
28. Wagner, B., Benecke, W., "Microfabricated actuator with moving permanent magnet," *Proceedings of the IEEE Conference on Micro Electro Mechanical Systems,* 27–32, 1991.
29. Yanagisawa, K., Tago, A., Ohkubo, T., Kuwano, H., "Magnetic microactuator," *Proceedings of the IEEE Conference on Micro Electro Mechanical Systems,* 120–124, 1991.
30. Judy, J., Muller, R. S., Zappe, H. H., "Magnetic microactuation of polysilicon flexure structures," *Journal of Microelectromechanical Systems,* 4(4):162–169, 1995.
31. Ahn, C. H., Kim, Y. J., Allen, M. G., "A planar variable reluctance magnetic micromotor with fully integrated stator and wrapped coils," *Proceedings of the IEEE Conference on Micro Electro Mechanical Systems,* 1–6, 1993.
32. Guckel, H., Christenson, T. R., Skrobis, K. J., Jung, T. S., Klein, J., Hartojo, K. V., Widjaja, I., "A first functional current excited planar rotational magnetic micromotor," *Proceedings of the IEEE Conference on Micro Electro Mechanical Systems,* 7–11, 1993.
33. Guckel, H., Sneigowski, J. J., Christenson, T. R., Raissi, F., "The application of fine grained, tensile polysilicon to mechanically resonant transducers," *Sensor and Actuators,* A21–A23:346–351, 1990.
34. Ikeda, K., Kuwayama, H., Kobayashi, T., Watanabe, T., Nishikawa, T., Yoshida, T., Harada, K., "Silicon pressure sensor integrates resonant strain gauge on diaphragm," *Sensors and Actuators,* A21–A23:146–150, 1990.
35. Fujii, T., Gotoh, Y., Kuroyanagi, S., "Fabrication of microdiaphragm pressure sensor utilizing micromachining," *Sensors and Actuators,* A34:217–224, 1992.
36. Mallon, J., Pourahmadi, F., Petersen, K., Barth, P., Vermeulen, T., Bryzek, J., "Low-pressure sensors employing bossed diaphragms and precision etch-stopping," *Sensors and Actuators,* A21–23:89–95, 1990.
37. Ikeda, K., Kuwayama, H., Kobayashi, T., Watanabe, T., Nishikawa, T., Yoshida, T., Harada, K., "Three-dimensional micromachining of silicon pressure sensor integrating resonant strain gauge on diaphragm," *Sensors and Actuators,* A21–A23:1007–1009, 1990.

38. Nagata, T., Terabe, H., Kuwahara, S., Sakurai, S., Tabata, O., Sugiyama, S., Esashi, M., "Digital compensated capacitive pressure sensor using cmos technology for low-pressure measurements," *Sensors and Actuators,* A34:173–177, 1992.
39. Stemme, E., Stemme, G., "A balanced resonant pressure sensor," *Sensors and Actuators,* A21–A23: 336–341, 1990.
40. Boxenhorn, B., Greiff, P., "Monolithic silicon accelerometer," *Sensors and Actuators,* A21–A23:273–277, 1990.
41. Leuthold, H., Rudolf, F., "An ASIC for high-resolution capacitive microaccelerometers," *Sensors and Actuators,* A21–A23:278–281, 1990.
42. Seidel, H., Riedel, H., Kolbeck, R., Muck, G., Kupke, W., Koniger, M., "Capacitive silicon accelerometer with highly symmetrical design," *Sensors and Actuators,* A21–A23:312–315, 1990.
43. Chang, S. C., Putty, M. W., Hicks, D. B., Li, C. H., Howe, R. T., "Resonant-bridge two-axis microaccelerometer," *Sensors and Actuators,* A21–A23:342–345, 1990.
44. Satchell, D. W., Greenwood, J. C., "A thermally-excited silicon accelerometer," *Sensors and Actuators,* A17:241–245, 1989.
45. Chau, K. H. L., Lewis, S. R., Zhao, Y., Howe, R. T., Bart, S. F., Marchesilli, R. G., "An integrated force- balanced capacitive accelerometer for low-*g* applications," *Sensors and Actuators,* A54:472–476, 1996.
46. Kuehnel, W., Sherman, S., "A surface micromachined silicon accelerometer with on-chip detection circuitry," *Sensors and Actuators,* A45:7–16, 1994.
47. Despont, Racine, G. A., Renaud, P., de Rooij, N. F., "New design of micromachined capacitive force sensor," *Journal of Micromechanics and Microengineering,* 3:239–242, 1993.
48. Blom, F. R., Bouwstra, S., Fluitman, J. H. J., Elwenspoek, M., "Resonating silicon beam force sensor," *Sensors and Actuators,* 17:513–519, 1989.
49. Bernstein, J., Cho, S., King, A. T., Kourepenis, A., Maciel, P., Weinberg, M., "A micromachined comb-drive tuning fork rate gyroscope," *IEEE Conference on Micro Electro Mechanical Systems,* 143–148, 1993.
50. Oh, Y., Lee, B., Baek, S., Kim, H., Kim, J., Kang, S., Song, C., "A surface-micromachined tunable vibratory gyroscope," *IEEE Conference on Micro Electro Mechanical Systems,* 272–277, 1997.
51. Venturi, M., Credi, A., Balzani, V., "Devices and machines at the molecular level," *Electronic Properties of Novel Materials, AIP Conf. Proc.,* 544:489–494, 2000.
52. Ajayan, P. M., Charlier, J. C., Rinzler, A. G., "PNAS," 96:14199–14200, 1999.
53. Zettl, A., Cumings, J., "Sharpened nanotubes, nanobearings and nanosprings," *Electronic Properties of Novel Materials, AIP Conf. Proc.,* 544:526–531, 2000.
54. Baughman, R. H., et al., "Carbon nanotube actuators," *Science,* 284:1340–1344, 1999.
55. Kim, P., Lieber, C. M., "Nanotube nanotweezers," *Science,* 286:2148–2150, 1999.
56. Freitas, R. A., "Nanomedicine," Vol. 1, *Landes Bioscience,* Austin, 1999.

6

Mechatronics: New Directions in Nano-, Micro-, and Mini-Scale Electromechanical Systems Design, and Engineering Curriculum Development

Sergey Edward Lyshevski

Purdue University Indianapolis

6.1 Introduction

Modern engineering encompasses diverse multidisciplinary areas. Therefore, there is a critical need to identify new directions in research and engineering education addressing, pursuing, and implementing new meaningful and pioneering research initiatives and designing the engineering curriculum. By integrating various disciplines and tools, mechatronics provides multidisciplinary leadership and supports the current gradual changes in academia and industry. There is a strong need for an advanced research in mechatronics and a curriculum reform for undergraduate and graduate programs. Recent research developments and drastic technological advances in electromechanical motion devices, power electronics, solid-state devices, microelectronics, micro- and nanoelectromechanical systems (MEMS and NEMS), materials and packaging, computers, informatics, system intelligence, microprocessors and

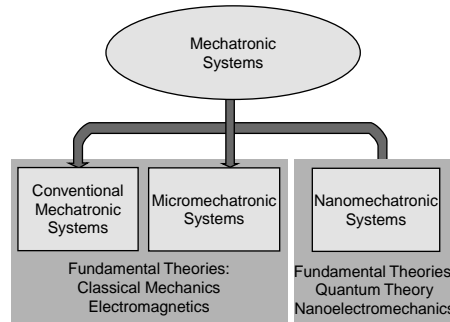



FIGURE 6.1 Classification and fundamental theories applied in mechatronic systems.

DSPs, signal and optical processing, computer-aided-design tools, and simulation environments have brought new challenges to the academia. As a result, many scientists are engaged in research in the area of mechatronics, and engineering schools have revised their curricula to offer the relevant courses in mechatronics.

Mechatronic systems are classified as:

1. conventional mechatronic systems,
2. microelectromechanical-micromechatronic systems (MEMS), and
3. nanoelectromechanical-nanomechatronic systems (NEMS).

The operational principles and basic foundations of conventional mechatronic systems and MEMS are the same, while NEMS can be studied using different concepts and theories. In particular, the designer applies the classical mechanics and electromagnetics to study conventional mechatronic systems and MEMS. Quantum theory and nanoelectromechanics are applied for NEMS, see Fig. 6.1.

One weakness of the computer, electrical, and mechanical engineering curricula is the well-known difficulties to achieving sufficient background, knowledge, depth, and breadth in integrative electromechanical systems areas to solve complex multidisciplinary engineering problems. Mechatronics introduces the subject matter, multidisciplinary areas, and disciplines (e.g., electrical, mechanical, and computer engineering) from unified perspectives through the electromechanical theory fundamentals (research) and designed sequence of mechatronic courses within an electromechanical systems (mechatronic) track or program (curriculum). This course sequence can be designed based upon the program objectives, strength, and goals. For different engineering programs (e.g., electrical, mechanical, computer, aerospace, material), the number of mechatronic courses, contents, and coverage are different because mechatronic courses complement the basic curriculum. However, the ultimate goal is the same: educate and prepare a new generation of students and engineers to solve a wide spectrum of engineering problems.

Mechatronics is an important part of modern confluent engineering due to integration, interaction, interpretation, relevance, and systematization features. Efficient and effective means to assess the current trends in modern engineering with assessments analysis and outcome prediction can be approached through the mechatronic paradigm. The multidisciplinary mechatronic research and educational activities, combined with the variety of active student learning processes and synergetic teaching styles, will produce a level of overall student accomplishments that is greater than the achievements which can be produced by refining the conventional electrical, computer, and mechanical engineering curricula. The multidisciplinary mechatronic paradigm serves very important purposes because it brings new depth to engineering areas, advances students' knowledge and background, provides students with the basic problem-solving skills that are needed to cope with advanced electromechanical systems controlled by microprocessors or DSPs, covers state-of-the-art hardware, and emphasizes and applies

modern software environments. Through the mechatronic curriculum, important program objectives and goals can be achieved. The integration of mechatronic courses into the engineering curriculum is reported in this chapter. Our ultimate goal is to identify the role, examine the existing courses, refine and enhance mechatronic curriculum in order to improve the structure and content of engineering programs, recruit and motivate students, increase teaching effectiveness and improve material delivery, as well as assess and evaluate the desired engineering program outcomes. The primary emphasis is placed on enhancement and improvement in student knowledge, learning, critical thinking, depth, breadth, results interpretation, integration and application of knowledge, motivation, commitment, creativity, enthusiasm, and confidence. These can be achieved through the mechatronic curriculum development and implementation. This chapter reports the development of a mechatronic curriculum. The role of mechatronics in modern engineering is discussed and documented.

6.2 Nano-, Micro-, and Mini-Scale Electromechanical Systems and Mechatronic Curriculum

Conventional, mini- and micro-scale electromechanical systems are studied from a unified perspective because operating features, basic phenomena, and dominant effects are based upon classical electromagnetics and mechanics (electromechanics). Electromechanical systems integrate subsystems and components. No matter how well an individual subsystem or component (electric motor, sensor, power amplifier, or DSP) performs, the overall performance can be degraded if the designer fails to integrate and optimize the electromechanical system. While electric machines, sensors, power electronics, microcontrollers, and DSPs should be emphasized, analyzed, designed, and optimized, the main focus is centered on integrated issues. The designer sometimes fails to grasp and understand the global picture because this requires extensive experience, background, knowledge, and capabilities to attain detailed assessment analysis with outcome prediction and overall performance evaluation. While the component-based *divide-and-solve* approach is valuable and applicable in the preliminary design phase, it is very important that the design and analysis of integrated electromechanical systems be accomplished in the context of global optimization with proper objectives, specifications, requirements, and bounds imposed. Novel electromechanical and VLSI technologies, computer-aided-design software, software-hardware co-design tools, high-performance software environments, and robust computational algorithms must be applied to design electromechanical systems. The main objective of the mechatronic curriculum development is to satisfy academia–industry–government demands as well as to help students develop in-depth fundamental, analytic, and experimental skills in analysis, design, optimization, control, and implementation of advanced integrated electromechanical systems. It is not possible to cover the full spectrum of mechatronics issues in a single course. Therefore, the mechatronic curriculum must be developed assuming that students already have sufficient fundamentals in calculus, physics, circuits, electromechanical devices, sensors, and controls.

The engineering curriculum usually integrates general education, science, and engineering courses. The incorporation of multidisciplinary engineering science and engineering design courses represents a major departure from the conventional curriculum. Usually, even electrical engineering students have some deficiencies in advanced electromagnetics, electric machinery, power electronics, ICs, microcontrollers, and DSPs because several of these courses are elective. Mechanical engineering students, while advancing electrical engineering students in mechanics and thermodynamics, have limited access to electromagnetics, electric machines, power electronics, microelectronics, and DSP courses. In addition, there are deficiencies in computer science and engineering mathematics for both electrical and mechanical engineering students because these courses are usually required only for computer engineering students. The need for engineering mathematics, electromagnetics, power electronics, and electromechanical motion devices (electric machines, actuators, and sensors) has not diminished, rather strengthened. In addition, radically new advanced hardware has been developed using enabling

fabrication technologies to fabricate nano- and micro-scale sensors, actuators, ICs, and antennas. Efficient software has emerged. To overcome the difficulties encountered, the mechatronic courses which cover the multidisciplinary areas must be introduced to the engineering curriculum. Mechatronics has been enthusiastically explored and supported by undergraduate and graduate, educational and research-oriented universities, high-technology industry, and government laboratories. However, there is a need to develop the long-term strategy in mechatronic research and education, define the role, as well as implement, commercialize, and market the mechatronic and electromechanics programs.

6.3 Mechatronics and Modern Engineering

Many engineering problems can be formulated, attacked, and solved using the mechatronic paradigm. Mechatronics deals with benchmarking and emerging problems in integrated electrical–mechanical–computer engineering, science, and technologies. Many of these problems have not been attacked and solved; and sometimes, the existing solutions cannot be treated as the optimal one. This reflects obvious trends in fundamental, applied, and experimental research as well as curriculum changes in response to long-standing unsolved problems, engineering and technological enterprise, and entreaties of steady evolutionary demands.

Mechatronics is the integrated design, analysis, optimization, and virtual prototyping of intelligent and high-performance electromechanical systems, system intelligence, learning, adaptation, decision making, and control through the use of advanced hardware (actuators, sensors, microprocessors, DSPs, power electronics, and ICs) and leading-edge software.

Integrated multidisciplinary features approach quickly, as documented in Fig. 6.2. The mechatronic paradigm, which integrates electrical, mechanical, and computer engineering, takes place.

The structural complexity of mechatronic systems has increased drastically due to hardware and software advancements, as well as stringent *achievable* performance requirements. Answering the demands of rising electromechanical system complexity, performance specifications, and intelligence, the mechatronic paradigm was introduced. In addition to the proper choice of electromechanical system components and subsystems, there are other issues which must be addressed in view of the constantly evolving nature of the electromechanical systems theory (e.g., analysis, design, modeling, optimization, complexity, intelligence, decision making, diagnostics, packaging). Competitive *optimum-performance* electromechanical systems must be designed within the advanced hardware and software concepts.

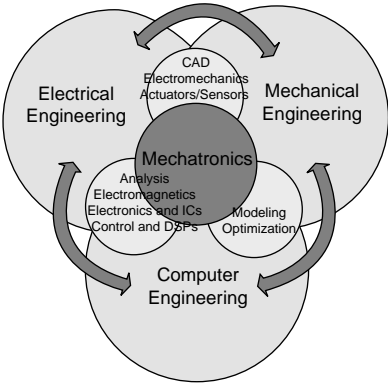



FIGURE 6.2 Mechatronics integrates electrical, mechanical, and computer engineering.

6.4 Design of Mechatronic Systems

One of the most challenging problems in mechatronic systems design is the system architecture synthesis, system integration, optimization, as well as selection of hardware (actuators, sensors, power electronics, ICs, microcontrollers, and DSPs) and software (environments, tools, computation algorithms to perform control, sensing, execution, emulation, information flow, data acquisition, simulation, visualization, virtual prototyping, and evaluation). Attempts to design state-of-the-art high-performance mechatronic systems and to guarantee the integrated design can be pursued through analysis of complex patterns and paradigms of evolutionary developed biological systems. Recent trends in engineering have increased the emphasis on integrated analysis, design, and control of advanced electromechanical systems. The scope of mechatronic systems has continued to expand, and, in addition to actuators, sensors, power electronics, ICs, antennas, microprocessors, DSPs, as well as input/output devices, many other subsystems must be integrated. The design process is evolutionary in nature. It starts with a given set of requirements and specifications. High-level functional design is performed first in order to produce detailed design at the subsystem and component level. Using the advanced subsystems and components, the initial design is performed, and the closed-loop electromechanical system performance is tested against the requirements. If requirements and specifications are not met, the designer revises or refines the system architecture, and other solutions are sought. At each level of the design hierarchy, the system performance in the behavioral domain is used to evaluate and refine the design process and solution devised. Each level of the design hierarchy corresponds to a particular abstraction level and has the specified set of activities and design tools that support the design at this level. For example, different criteria are used to design actuators and ICs due to different behavior, physical properties, operational principles, and performance criteria imposed for these components. It should be emphasized that the level of hierarchy must be defined, e.g., there is no need to study the behavior of millions of transistors on each IC chip because mechatronic systems integrate hundreds of ICs, and the end-to-end behavior of ICs is usually evaluated (ICs are assumed to be optimized, and these ICs are used as ready-to-use components). The design flow is illustrated in Fig. 6.3.

Automated synthesis can be attained to implement this design flow. The design of mechatronic systems is a process that starts from the specification of requirements and progressively proceeds to perform a functional design and optimization that is gradually refined through a sequence of steps. Specifications typically include the performance requirements derived from systems functionality, operating envelope, affordability, and other requirements. Both *top-down* and *bottom-up* approaches should be combined to design high-performance mechatronic systems augmenting hierarchy, integrity, regularity, modularity, compliance, and completeness in the synthesis process. Even though the

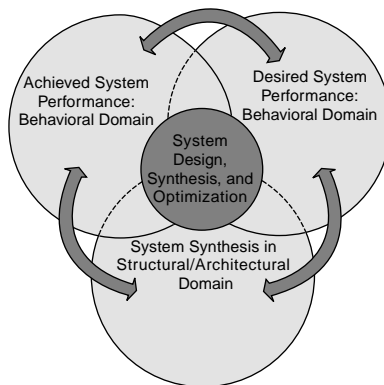



FIGURE 6.3 Design flow in synthesis of mechatronic systems.

basic foundations have been developed, some urgent areas have been downgraded, less emphasized, and researched. The mechatronic systems synthesis reported guarantees an eventual consensus between behavioral and structural domains, as well as ensures descriptive and integrative features in the design. These were achieved applying the mechatronic paradigm which allows one to extend and augment the results of classical mechanics, electromagnetics, electric machinery, power electronics, microelectronics, informatics, and control theories, as well as to apply advanced integrated hardware and software.

To acquire and expand the engineering core, there is the need to augment interdisciplinary areas as well as to link and place the multidisciplinary perspectives integrating actuators–sensors–power electronics–ICs–DSPs to attain actuation, sensing, control, decision making, intelligence, signal processing, and data acquisition. New developments are needed. The theory and engineering practice of high-performance electromechanical systems should be considered as the unified cornerstone of the engineering curriculum through mechatronics. The unified analysis of actuators and sensors (e.g., electromechanical motion devices), power electronics and ICs, microprocessors and DSPs, and advanced hardware and software, have barely been introduced into the engineering curriculum. Mechatronics, as the breakthrough concept in the design and analysis of conventional-, mini-, micro- and nano-scale electromechanical systems, was introduced to attack, integrate, and solve a great variety of emerging problems.

6.5 Mechatronic System Components

Mechatronics integrates electromechanical systems design, modeling, simulation, analysis, software-hardware developments and co-design, intelligence, decision making, advanced control (including self-adaptive, robust, and intelligent motion control), signal/image processing, and virtual prototyping. The mechatronic paradigm utilizes the fundamentals of electrical, mechanical, and computer engineering with the ultimate objective to guarantee the synergistic combination of precision engineering, electronic control, and intelligence in the design, analysis, and optimization of electromechanical systems. Electromechanical systems (robots, electric drives, servomechanisms, pointing systems, assemblers) are highly nonlinear systems, and their accurate actuation, sensing, and control are very challenging problems. Actuators and sensors must be designed and integrated with the corresponding power electronic subsystems. The principles of matching and compliance are general design principles, which require that the electromechanical system architectures should be synthesized integrating all subsystems and components. The matching conditions have to be determined and guaranteed, and actuators– sensors–power electronics compliance must be satisfied. Electromechanical systems must be controlled, and controllers should be designed. Robust, adaptive, and intelligent control laws must be designed, examined, verified, and implemented. The research in control of electromechanical systems aims to find methods for devising intelligent and motion controllers, system architecture synthesis, deriving feedback maps, and obtaining gains. To implement these controllers, microprocessors and DSPs with ICs (input-output devices, A/D and D/A converters, optocouplers, transistor drivers) must be used. Other problems are to design, optimize, and verify the analysis, control, execution, emulation, and evaluation software.

It was emphasized that the design of high-performance mechatronic systems implies the subsystems and components developments. One of the major components of mechatronic systems are electric machines used as actuators and sensors. The following problems are usually emphasized: characterization of electric machines, actuators, and sensors according to their applications and overall systems requirements by means of specific computer-aided-design software; design of high-performance electric machines, actuators, and sensors for specific applications; integration of electric motors and actuators with sensors, power electronics, and ICs; control and diagnostic of electric machines, actuators, and sensors using microprocessors and DSPs.

6.6 Systems Synthesis, Mechatronics Software, and Simulation

Modeling, simulation, and synthesis are complementary activities performed in the design of mechatronic systems. Simulation starts with the model developments, while synthesis starts with the specifications imposed on the behavior and analysis of the system performance through analysis using modeling, simulation, and experimental results. The designer mimics, studies, analyzes, and evaluates the mechatronic system's behavior using state, performance, control, events, disturbance, and other variables. The synthesis process was described in section 6.4. Modeling, simulation, analysis, virtual prototyping, and visualization are critical and urgently important aspects for developing and prototyping of advanced electromechanical systems. As a flexible high-performance modeling and design environment, MATLAB has become a standard, cost-effective tool. Competition has prompted cost and product cycle reductions. To speed up analysis and design with assessment analysis, facilitate enormous gains in productivity and creativity, integrate control and signal processing using advanced microprocessors and DSPs, accelerate prototyping features, generate real-time C code and visualize the results, perform data acquisition and data intensive analysis, the MATLAB® environment is used. In MATLAB, the following commonly used toolboxes can be applied: SIMULINK®, Real-Time Workshop™, Control System, Nonlinear Control Design, Optimization, Robust Control, Signal Processing, Symbolic Math, System Identification, Partial Differential Equations, Neural Networks, as well as other application-specific toolboxes (see the MATLAB demo typing `demo` in the Command Window). MATLAB capabilities should be demonstrated by attacking important practical examples in order to increase students' productivity and creativity by demonstrating how to use the advanced software in electromechanical system applications. The MATLAB environment offers a rich set of capabilities to efficiently solve a variety of complex analysis, modeling, simulation, control, and optimization problems encountered in undergraduate and graduate mechatronic courses. A wide array of mechatronic systems can be modeled, simulated, analyzed, and optimized. The electromechanical systems examples, integrated within mechatronic courses, will provide the practice and educate students with the highest degree of comprehensiveness and coverage.

6.7 Mechatronic Curriculum

The ultimate objective of the mechatronic curriculum is to educate a new generation of students and engineers, as well as to assist industry and government in the development of high-performance electromechanical systems augmenting conventional engineering curriculum with an ever-expanding electromechanics core. The emphasis should be focused on advancing the overall mission of the engineering curriculum, because through mechatronics it is possible to further define, refine, and expand the objectives into three fundamental areas, which are research, education, and service. Using the mechatronic paradigm, academia will perform world-class fundamental and applied research by

- integrating electromagnetics, electromechanics, power electronics, ICs, and control;
- devising advanced design, analysis, and optimization simulation and analytic tools and capabilities through development of specialized computer-aided-design software;
- developing actuation-sensing-control hardware;
- devising advanced paradigms, concepts, and technologies;
- supporting research, internship, and cooperative multidisciplinary education programs for undergraduate and graduate students;
- supporting, sustaining, and assisting faculty in emerging new areas.

Mechatronic curriculum design includes development of goals and objectives, programs of study and curriculum guides, courses, laboratories, textbooks, instructional materials, manuals, experiments,

instructional sequences, material delivery techniques, visualization and demonstration approaches, and other supplemental materials to accomplish a wide range of educational and research goals. There is an increase in the number of students whose good programming skills and theoretical background match with complete inability to solve simple engineering problems. The fundamental goal of mechatronic courses is to demonstrate the application of theoretical, applied, and experimental results in analysis, design, and deployment of complex electromechanical systems (including NEMS and MEMS), to cover emerging hardware and software, to introduce and deliver the rigorous theory of electromechanics, to help students develop strong problem-solving skills, as well as to provide the needed engineering practice. The courses in mechatronics are intended to develop a thorough understanding of integrated perspectives in analysis, modeling, simulation, optimization, design, and implementation of complex electromechanical systems. By means of practical, worked-out examples, students will be prepared and trained to use the results in engineering practice, research, and developments. Advanced hardware and software of engineering importance (electromechanical motion devices, actuators, sensors, solid-state devices, power electronics, ICs, microprocessors, and DSPs) must be comprehensively covered in detail from multidisciplinary integrated perspectives.

At Purdue University Indianapolis, in the Department of Electrical and Computer Engineering, the following undergraduate courses are required in the Electrical Engineering plan of study: *Linear Circuit Analysis* I and II, *Signals and Systems, Semiconductor Devices, Electric and Magnetic Fields, Microprocessor Systems and Interfacing*, and *Feedback Systems Analysis and Design*. The following elective undergraduate courses assist the mechatronic area: *Electromechanical Motion Devices, Computer Architecture, Digital Signal Processing*, and *Multimedia Systems*. In addition to this set of core Electrical and Computer Engineering courses, there is a critical need to teach the courses in mechatronics.

The mechatronic curriculum should emphasize and augment traditional engineering topics and the latest enabling technologies and developments to integrate and stimulate new advances in the analysis and design of advanced state-of-the-art mechatronic systems. For example, the following courses should be developed and offered: *Mechatronic Systems, Smart Structures, Micromechatronics* (*Microelectromechanical Systems*), and *Nanomechatronics* (*Nanoelectromechanical Systems*).

The major goal is to ensure a deep understanding of the engineering underpinnings, integrate engineering–science–technology, and develop the modern picture of electromechanical engineering by using the bedrock fundamentals of mechatronics. It is recognized by academia, industry, and government that the most urgent areas of modern mechatronics needing development are MEMS and NEMS. Therefore, current developments should be concentrated to perform fundamental, applied, and experimental research in these emerging fields.

6.8 Introductory Mechatronic Course

At Purdue University Indianapolis, in the Electrical and Computer Engineering and Mechanical Engineering departments, an Electrical/Mechanical Engineering senior-level undergraduate–junior graduate mechatronic course was developed and offered. The topics covered are given in Table 6.1.

This course is developed to bridge the engineering–science–technology gap by bonding innovative multi-disciplinary developments, focusing on state-of-the-art hardware, and centering on high-performance software. The developed course dramatically reduces the time students need to establish basic skills for high-technology employability. The objective of this course is twofold: to bring recent developments of modern electromechanics and to integrate an interactive studio-based method of instruction and delivery. During the past decade, there has been a shift in engineering education from an instructor-centered lectures environment to a student-centered learning environment. We have developed a mechatronics studio that combines lectures, simulation exercises, and experiments in a single classroom in order to implement new teaching and delivery methods through an active learning environment, activity-based strategies, interactive multimedia, networked computer-based learning, multisynchronous delivery of supporting materials, and effective demonstration. Simulation-based assignments can be used to illustrate problems that cannot be easily studied and assessed using classical paper-and-pencil analytic solutions.

TABLE 6.1 Mechatronic Course Contents

No.	Topic	Class
1	Introduction to electromechanical systems and mechatronics	1
2	Electromagnetics and mechanics in mechatronic systems: Newtonian mechanics, the Lagrange equations of motion, and Kirchhoff's laws	2
3	Energy conversion and electromechanical analogies	2
4	Dynamics of mechatronic system	2
5	The MATLAB environment in nonlinear analysis and modeling of mechatronic systems	2
6	Permanent-magnet direct-current and synchronous servo-motors	4
7	Transducers and smart structures: actuators and sensors	2
8	Power electronics, driving circuitry, power converters and amplifiers	4
9	Motion control of electromechanical systems and smart structures	3
10	Microprocessors and DSPs in control and data acquisition of mechatronic systems	2
11	Mechatronic systems: case-studies, modeling, analysis, control, and laboratory experiments	3
12	Advanced project	1

Although simulation-based assignments provide much insight to practical problems, there is nothing that can take the place of hands-on experiments. The mechatronics is introduced through synergy of comprehensive systems design, high-fidelity modeling, simulation, hardware demonstration, and case studies.

The assessment performed demonstrates that this course guarantees comprehensive, balanced coverage, satisfies the program objectives, and fulfills the goals. While students are familiar with some topics of advanced engineering and science (calculus and physics), it is clear that they do not have sufficient background in nonlinear dynamics and control, electric machinery, power electronics, solid-state devices, ICs, microprocessors, and DSPs. Therefore, the material is presented in sufficient details, and basic theory needed to fully understand, appreciate, and apply mechatronics is covered. In this course, most efficient and straightforward analysis, modeling, simulation, and synthesis methods are presented and demonstrated with ultimate objectives to address and solve the analysis, design, control, optimization, and virtual prototyping problems. A wide range of worked-out examples and qualitative illustrations, which are treated in-depth, bridge the gap between the theory, practical problems, and engineering practice. Step-by-step, the mechatronic course guides students from rigorous theoretical foundation to advanced applications and implementation. In addition to achieving a good balance between theory and application, state-of-the-art hardware and software are emphasized and demonstrated. In this course, mechatronic systems are thoroughly covered, and students can easily apply the results to attack real engineering problems.

6.9 Books in Mechatronics

The demand for educational books in mechatronics far exceeds what was previously anticipated by academia and industry. Excellent textbooks in electric machinery [1–8], power electronics [9–11], microelectronics and ICs [12], and sensors [13,14] were published. Educational examples in analysis and design of linear electromechanical systems are available from control books [15–21]. *Control Systems Theory With Engineering Applications* [18], shown in Fig. 6.4, has a number of illustrative examples in modeling, simulation, and control of complex nonlinear electromechanical systems. In particular, analysis and control of nonlinear transducers, permanent-magnet DC and synchronous motors, squirrel-cage induction motors, servomechanisms, and power converters are thoroughly covered.

The need for a comprehensive treatment of nonlinear electromechanical systems using the mechatronic paradigm is evident. Excellent books in conventional electromechanical motion devices [3,4,22], and textbooks for mechanical engineering students in mechatronics [23–27] have been used in Electrical and Mechanical Engineering departments, respectively. However, there is a critical need for modern books in mechatronics that are comprehensive in their coverage and global in their perspective for engineering departments. The time has come to target new frontiers using the developed engineering enterprise,

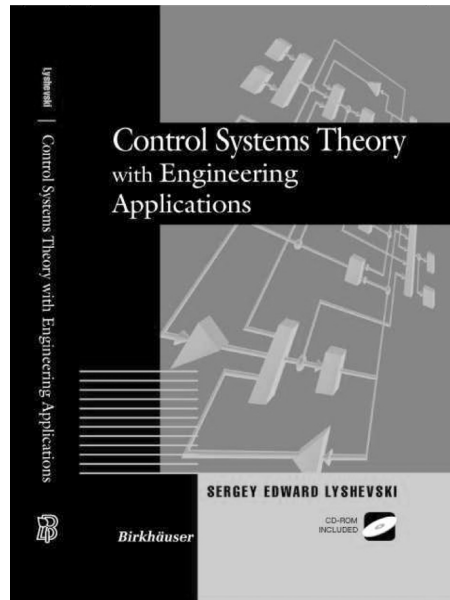



FIGURE 6.4 Control book with coverage in analysis and control of electromechanical systems. http://www.birkhauser.com/cgi-win/ISBN/0-8176-4203-X.

http://www.crcpress.com/us/product.asp?sku=2275&dept%5Fid=1 http://www.crcpress.com/us/product.asp?sku=0916&dept%5Fid=1

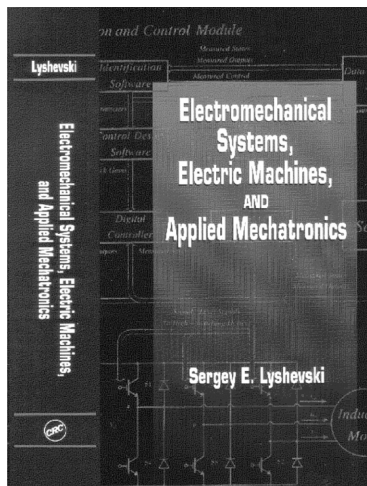
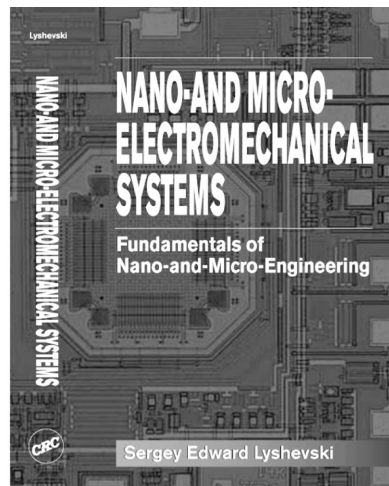



FIGURE 6.5 Books in electromechanical and mechatronic systems.

emerging technologies, advanced hardware, and state-of-the-art software. The book *Electromechanical Systems, Electric Machines, and Applied Mechatronics* [28] was written by taking advantage of the modern engineering curriculum, see Fig. 6.5. In this book, the fundamental theory of electromechanics, new enabling technologies, basic engineering principles, system integration, modeling, analysis, simulation, control, as well as a spectrum of emerging engineering problems, were comprehensively covered. For NEMS and MEMS, the book *Nano- and Micro-Electromechanical Systems: Fundamentals of Nano- and Micro-Engineering* [29] can be effectively used. A wide number of demonstrations and examples of electromechanical systems are covered.

6.10 Mechatronic Curriculum Developments

The current mechatronic curriculum leaves much to be desired, and the following strategy, which can be modified and expanded, should be pursued by academia to integrate the mechatronic courses in the undergraduate and graduate curricula:

- commercialize and market mechatronic program;
- expand the mechatronic horizon to conventional and mini-scale mechatronic systems, as well as to MEMS and NEMS which are emerging areas in engineering;
- revise the engineering curriculum. In particular, Electromagnetics, Electromechanical Motion Devices, Power Electronics, Control, Microelectronics, and DSP courses should be offered as the required core courses, and as prerequisites for advanced mechatronic courses;
- emphasize mechatronics as the center of the undergraduate and graduate electromechanical engineering curriculum rather than at the periphery;
- cover moderately complex electromechanical systems and case studies in the undergraduate mechatronic courses and relocate highly specialized topics to the graduate program;
- develop an intellectually demanding, progressive, well-balanced mechatronic curriculum and mechatronic courses with laboratories;
- fully integrate computer-aided-design tools and advanced high-performance simulation software;
- extend mechatronics to the undergraduate senior design projects;
- write and publish comprehensive books, textbooks, and handbooks in mechatronics; and
- widely and timely disseminate the results.

Manageable collaboration between engineering disciplines and departments can be achieved within the mechatronic program. The following basic courses sequence can be applied:

- Electromechanical Motion Devices,
- Power Electronics and Microelectronics,
- Microprocessors and Interfacing,
- Digital Signal Processing,
- Electromechanical Systems,
- Introduction to Mechatronics,
- Control Systems Theory and Control of Mechatronic Systems,
- Mechatronic Systems and Smart Structures,
- Microelectromechancial Systems,
- Nanoelectromechanical Systems.

Due to the differences in the electrical and computer, mechanical, and aerospace engineering plans of study and the limited number of elective engineering courses counted towards the degree, the mechatronic courses sequence can be different. For example, for electrical engineering students, the coursework plan of study can be designed using fundamental electrical engineering and applied mechanical engineering; for mechanical engineering students, fundamental mechanical engineering and applied electrical engineering can be emphasized. The students will have fundamentals in one core area while accomplishing breadth and receiving applied knowledge in the other field.

6.11 Conclusions: Mechatronics Perspectives

Far-reaching fundamental and technological advances in electromechanical motion devices (actuators and sensors), power electronics, solid-state devices, ICs, MEMS and NEMS, materials and packaging, computers and informatics, microprocessors and DSPs, digital signal and optical processing, as well as computer-aided-design tools and simulation software, have brought new challenges to academia,

industry, and government. As a result, many engineering schools have revised their curricula in order to offer the relevant interdisciplinary courses such as Electromechanical Systems and Mechatronics. The basis of mechatronics is fundamental theory and engineering practice. The attempts to introduce mechatronics have been only partially successful due to the absence of a long-term strategy. Therefore, coordinated efforts are sought. Most engineering curricula provide a single elective course to introduce mechatronics to electrical, computer, mechanical, and aerospace engineering students. Due to the lack of time, it is impossible to comprehensively cover the material and thoroughly emphasize the cross-disciplinary nature of mechatronics in one introductory course. As a result, this undergraduate or dual-level course might not adequately serve the students' professional needs and goals, and does not satisfy growing academia, industrial, and government demands. A set of core mechatronic courses should be integrated into the engineering curriculum, and laboratory- and project-oriented courses should be developed to teach and demonstrate advanced hardware and software with application to complex electromechanical systems. The relevance of fundamental theory, applied results, and experiments is very important and must be emphasized. The great power and versatility of mechatronics, not to mention the prime importance of the results it approaches in all areas of engineering, make it worthwhile for all engineers to be acquainted with the basic theory and engineering practice. There is no end to the application of mechatronics and to the further contribution to this interdisciplinary concept. We have just skimmed the surface of mechatronics application to advanced electromechanical systems. New trends will be researched and applied in the near future because mechatronics is an engineering–science–technology frontier. For example, novel phenomena and operating principles in NEMS and MEMS can be devised, studied, analyzed, and verified using nanomechatronics and nanoelectromechanics.

References

1. Chapman, S. J., *Electric Machinery Fundamentals,* McGraw-Hill, New York, 1999.
2. Fitzgerald, A. E., Kingsley, C., and Umans, S. D., *Electric Machinery,* McGraw-Hill, New York, 1990.
3. Krause, P. C., and Wasynczuk, O., *Electromechanical Motion Devices,* McGraw-Hill, New York, 1989.
4. Krause, P. C., Wasynczuk, O., and Sudhoff, S. D., *Analysis of Electric Machinery,* IEEE Press, New York, 1995.
5. Leonhard, W., *Control of Electrical Drives,* Springer, Berlin, 1996.
6. Ong, C. M., *Dynamic Simulation of Electric Machines,* Prentice-Hall, Upper Saddle River, NJ, 1998.
7. Novotny, D. W., and Lipo, T. A., *Vector Control and Dynamics of AC Drives,* Clarendon Press, Oxford, 1996.
8. Slemon, G. R., *Electric Machines and Drives,* Addison-Wesley Publishing Company, Reading, MA, 1992.
9. Hart, D. W., *Introduction to Power Electronics,* Prentice-Hall, Upper Saddle River, NJ, 1997.
10. Kassakian, J. G., Schlecht, M. F., and Verghese, G. C., *Principles of Power Electronics,* Addison-Wesley Publishing Company, Reading, MA, 1991.
11. Mohan, N. T., Undeland, M., and Robbins, W. P., *Power Electronics: Converters, Applications, and Design,* John Wiley and Sons, New York, 1995.
12. Sedra, A. S., and Smith, K. C., *Microelectronic Circuits,* Oxford University Press, New York, 1997.
13. Fraden, J., *Handbook of Modern Sensors: Physics, Design, and Applications,* AIP Press, Woodbury, NY, 1997.
14. Kovacs, G. T. A., *Micromachined Transducers Sourcebook,* McGraw-Hill, New York, 1998.
15. Dorf, R. C., and Bishop, R. H., *Modern Control Systems,* Addison-Wesley Publishing Company, Reading, MA, 1995.
16. Franklin, J. F., Powell, J. D., and Emami-Naeini, A., *Feedback Control of Dynamic Systems,* Addison-Wesley Publishing Company, Reading, MA, 1994.
17. Kuo, B. C., *Automatic Control Systems,* Prentice-Hall, Englewood Cliffs, NJ, 1995.
18. Lyshevski, S. E., *Control Systems Theory With Engineering Applications,* Birkhäuser, Boston, MA, 2001. http://www.birkhauser.com/cgi-win/ISBN/0-8176-4203-X

19. Ogata, K., *Discrete-Time Control Systems,* Prentice-Hall, Upper Saddle River, NJ, 1995.
20. Ogata, K., *Modern Control Engineering,* Prentice-Hall, Upper Saddle River, NJ, 1997.
21. Phillips, C. L., and Harbor, R. D., *Feedback Control Systems,* Prentice-Hall, Englewood Cliffs, NJ, 1996.
22. White, D. C., and Woodson, H. H., *Electromechanical Energy Conversion,* Wiley, New York, 1959.
23. Auslander, D. M., and Kempf, C. J., *Mechatronics: Mechanical System Interfacing,* Prentice-Hall, Upper Saddle River, NJ, 1996.
24. Bolton, W., *Mechatronics: Electronic Control Systems in Mechanical Engineering,* Addison-Wesley Logman Publishing, New York, 1999.
25. Bradley, D. A., Dawson, D., Burd, N. C., and Loader, A. J., *Mechatronics,* Chapman and Hall, New York, 1996.
26. Fraser, C., and Milne, J., *Electro-Mechanical Engineering,* IEEE Press, New York, 1994.
27. Shetty, D., and Kolk, R. A., *Mechatronics System Design,* PWS Publishing Company, New York, 1997.
28. Lyshevski, S. E., *Electromechanical Systems, Electric Machines, and Applied Mechatronics,* CRC Press, Boca Raton, FL, 1999. http://www.crcpress.com/us/product.asp?sku=2275&dept%5Fid=1
29. Lyshevski, S. E., *Nano- and Microelectromechanical Systems: Fundamentals of Nano- and Microengineering,* CRC Press, Boca Raton, FL, 2000. http://www.crcpress.com/us/product.asp?sku=0916&dept%5Fid=1

Physical System Modeling

7

Modeling Electromechanical Systems

Francis C. Moon
Cornell University

7.1 Introduction

Mechatronics describes the integration of mechanical, electromagnetic, and computer elements to produce devices and systems that monitor and control machine and structural systems. Examples include familiar consumer machines such as VCRs, automatic cameras, automobile air bags, and cruise control devices. A distinguishing feature of modern mechatronic devices compared to earlier controlled machines is the miniaturization of electronic information processing equipment. Increasingly computer and electronic sensors and actuators can be embedded in the structures and machines. This has led to the need for integration of mechanical and electrical design. This is true not only for sensing and signal processing but also for actuator design. In human size devices, more powerful magnetic materials and superconductors have led to the replacement of hydraulic and pneumatic actuators with servo motors, linear motors, and other electromagnetic actuators. At the material scale and in microelectromechanical systems (MEMS), electric charge force actuators, piezoelectric actuators, and ferroelectric actuators have made great strides.

While the materials used in electromechanical design are often new, the basic dynamic principles of Newton and Maxwell still apply. In spatially extended systems one must solve continuum problems using the theory of elasticity and the partial differential equations of electromagnetic field theory. For many applications, however, it is sufficient to use lumped parameter modeling based on i) rigid body dynamics

for inertial components, ii) Kirchhoff circuit laws for current-charge components, and iii) magnet circuit laws for magnetic flux devices.

In this chapter we will examine the basic modeling assumptions for inertial, electric, and magnetic circuits, which are typical of mechatronic systems, and will summarize the dynamic principles and interactions between the mechanical motion, circuit, and magnetic state variables. We will also illustrate these principles with a few examples as well as provide some bibliography to more advanced references in electromechanics.

7.2 Models for Electromechanical Systems

The fundamental equations of motion for physical continua are partial differential equations (PDEs), which describe dynamic behavior in both time and space. For example, the motions of strings, elastic beams and plates, fluid flow around and through bodies, as well as magnetic and electric fields require both spatial and temporal information. These equations include those of elasticity, elastodynamics, the Navier–Stokes equations of fluid mechanics, and the Maxwell–Faraday equations of electromagnetics. Electromagnetic field problems may be found in Jackson (1968). Coupled field problems in electric fields and fluids may be found in Melcher (1980) and problems in magnetic fields and elastic structures may be found in the monograph by Moon (1984). This short article will only treat solid systems.

Many practical electromechanical devices can be modeled by lumped physical elements such as mass or inductance. The equations of motion are then integral forms of the basic PDEs and result in coupled ordinary differential equations (ODEs). This methodology will be explored in this chapter. Where physical problems have spatial distributions, one can often separate the problem into spatial and temporal parts called *separation of variables*. The spatial description is represented by a finite number of spatial or eigenmodes each of which has its modal amplitude. This method again results in a set of ODEs. Often these coupled equations can be understood in the context of simple lumped mechanical masses and electric and magnetic circuits.

7.3 Rigid Body Models

Kinematics of Rigid Bodies

Kinematics is the description of motion in terms of position vectors **r**, velocities **v**, acceleration **a**, rotation rate vector ω, and generalized coordinates $\{q_k(t)\}$ such as relative angular positions of one part to another in a machine (Fig. 7.1). In a rigid body one generally specifies the position vector of one point, such as the center of mass $\mathbf{r}_c$, and the velocity of that point, say $\mathbf{v}_c$. The angular position of a rigid body is specified by angle sets call Euler angles. For example, in vehicles there are pitch, roll, and yaw angles (see, e.g., Moon, 1999). The angular velocity vector of a rigid body is denoted by ω. The velocity of a point in a rigid body other than the center of mass, $\mathbf{r}_p = \mathbf{r}_c + \rho$, is given by

$$\mathbf{v}_P = \mathbf{v}_c + \omega \times \rho \tag{7.1}$$

where the second term is a vector cross product. The angular velocity vector ω is a property of the entire rigid body. In general a rigid body, such as a satellite, has six degrees of freedom. But when machine elements are modeled as a rigid body, kinematic constraints often limit the number of degrees of freedom.

Constraints and Generalized Coordinates

Machines are often collections of rigid body elements in which each component is constrained to have one degree of freedom relative to each of its neighbors. For example, in a multi-link robot arm shown in Fig. 7.2, each rigid link has a revolute degree of freedom. The degrees of freedom of each rigid link are constrained by bearings, guides, and gearing to have one type of relative motion. Thus, it is convenient

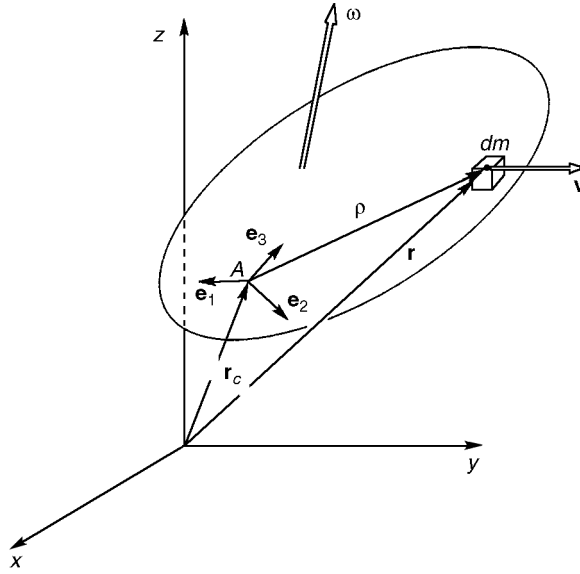



FIGURE 7.1 Sketch of a rigid body with position vector, velocity, and angular velocity vectors.

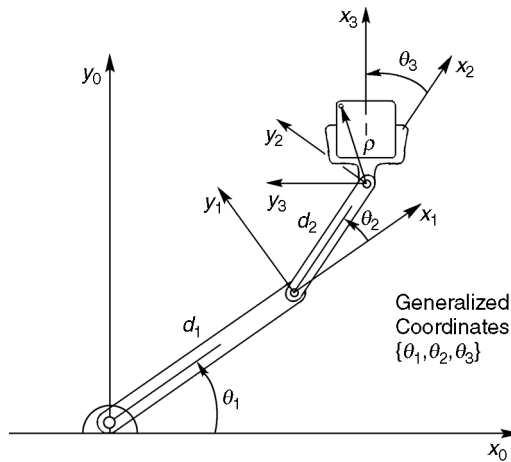



FIGURE 7.2 Multiple link robot manipulator arm.

to use these generalized motions $\{q_k\text{: } k = 1,\ldots,K\}$ to describe the dynamics. It is sometimes useful to define a vector or matrix, $\mathbf{J}(q_k)$, called a *Jacobian*, that relates velocities of physical points in the machine to the generalized velocities $\{\dot{q}_k\}$. If the position vector to some point in the machine is $\mathbf{r}_P(q_k)$ and is determined by geometric constraints indicated by the functional dependence on the $\{q_k(t)\}$, then the velocity of that point is given by

$$\mathbf{v}_P = \sum \frac{\partial \mathbf{r}_P}{\partial q_r} \dot{q}_r = \mathbf{J} \cdot \dot{\mathbf{q}} \tag{7.2}$$

where the sum is on the number of generalized degrees of freedom K. The three-by-K matrix $\mathbf{J}$ is called a *Jacobian* and $\dot{\mathbf{q}}$ is a $K \times 1$ vector of generalized coordinates. This expression can be used to calculate

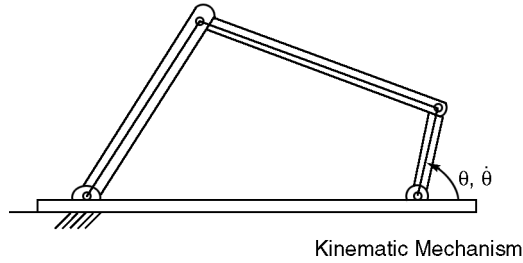



FIGURE 7.3 Example of a kinematic mechanism.

the kinetic energy of the constrained machine elements, and using Lagrange's equations discussed below, derive the equations of motion (see also Moon, 1999).

Kinematic versus Dynamic Problems

Some machines are constructed in a closed kinematic chain so that the motion of one link determines the motion of the rest of the rigid bodies in the chain, as in the four-bar linkage shown in Fig. 7.3. In these problems the designer does not have to solve differential equations of motion. Newton's laws are used to determine forces in the machine, but the motions are *kinematic*, determined through the geometric constraints.

In open link problems, such as robotic devices (Fig. 7.2), the motion of one link does not determine the dynamics of the rest. The motions of these devices are inherently *dynamic*. The engineer must use both the kinematic constraints (7.2) as well as the Newton–Euler differential equation of motion or equivalent forms such as Lagrange's equation discussed below.

7.4 Basic Equations of Dynamics of Rigid Bodies

In this section we review the equations of motion for the mechanical plant in a mechatronics system. This plant could be a system of rigid bodies such as in a serial robot manipulator arm (Fig. 7.2) or a magnetically levitated vehicle (Fig. 7.4), or flexible structures in a MEMS accelerometer. The dynamics of flexible structural systems are described by PDEs of motion. The equation for rigid bodies involves Newton's law for the motion of the center of mass and Euler's extension of Newton's laws to the angular momentum of the rigid body. These equations can be formulated in many ways (see Moon, 1999):

1. Newton–Euler equation (vector method)
2. Lagrange's equation (scalar-energy method)
3. D'Alembert's principle (virtual work method)
4. Virtual power principle (Kane's equation, or Jourdan's principle)

Newton–Euler Equation

Consider the rigid body in Fig. 7.1 whose center of mass is measured by the vector $\mathbf{r}_c$ in some fixed coordinate system. The velocity and acceleration of the center of mass are given by

$$\dot{\mathbf{r}}_c = \mathbf{v}_c, \qquad \dot{\mathbf{v}}_c = \mathbf{a}_c \tag{7.3}$$

The "over dot" represents a total derivative with respect to time. We represent the total sum of vector forces on the body from both mechanical and electromagnetic sources by **F**. Newton's law for the motion

FIGURE 7.4 Magnetically levitated rigid body (HSST MagLev prototype vehicle, 1998, Nagoya, Japan).

of the center of mass of a body with mass m is given by

$$m\dot{\mathbf{v}}_c = \mathbf{F} \tag{7.4}$$

If $\mathbf{r}$ is a vector to some point in the rigid body, we define a local position vector ρ by $\mathbf{r}_P = \mathbf{r}_c + \rho$. If a force $\mathbf{F}_i$ acts at a point $\mathbf{r}_i$ in a rigid body, then we define the moment of the force $\mathbf{M}$ about the fixed origin by

$$\mathbf{M}_i = \mathbf{r}_i \times \mathbf{F}_i \tag{7.5}$$

The total force moment is then given by the sum over all the applied forces as the body

$$\mathbf{M} = \sum \mathbf{r}_i \times \mathbf{F}_i = \mathbf{r}_c \times \mathbf{F} + \mathbf{M}_c \quad \text{where} \quad \mathbf{M}_c = \sum \rho_i \times \mathbf{F}_i \tag{7.6}$$

We also define the *angular momentum* of the rigid body by the product of a symmetric matrix of second moments of mass called the *inertia matrix* $\mathbf{I}_c$. The angular momentum vector about the center of mass is defined by

$$\mathbf{H}_c = \mathbf{I}_c \cdot \omega \tag{7.7}$$

Since $\mathbf{I}_c$ is a symmetric matrix, it can be diagonalized with principal inertias (or eigenvalues) $\{I_{ic}\}$ about principal directions (eigenvectors) $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$. In these coordinates, which are attached to the body, the angular momentum about the center of mass becomes

$$\mathbf{H}_c = I_{1c}\omega_1\mathbf{e}_1 + I_{2c}\omega_2\mathbf{e}_2 + I_{3c}\omega_3\mathbf{e}_3 \tag{7.8}$$

where the angular velocity vector is written in terms of principal eigenvectors $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ attached to the rigid body.

Euler's extension of Newton's law for a rigid body is then given by

$$\dot{\mathbf{H}}_c = \mathbf{M}_c \tag{7.9}$$

This equation says that the change in the angular momentum about the center of mass is equal to the total moment of all the forces about the center of mass. The equation can also be applied about a fixed point of rotation, which is not necessarily the center of mass, as in the example of the compound pendulum given below.

Equations (7.4) and (7.9) are known as the Newton–Euler equations of motion. Without constraints, they represent six coupled second order differential equations for the position of the center of mass and for the angular orientation of the rigid body.

Multibody Dynamics

In a serial link robot arm, as shown in Fig. 7.2, we have a set of connected rigid bodies. Each body is subject to both applied and constraint forces and moments. The dynamical equations of motion involve the solution of the Newton–Euler equations for each rigid link subject to the geometric or kinematics constraints between each of the bodies as in (7.2). The forces on each body will have applied terms $\mathbf{F}^a$, from actuators or external mechanical sources, and internal constraint forces $\mathbf{F}^c$. When friction is absent, the work done by these constraint forces is zero. This property can be used to write equations of motion in terms of scalar energy functions, known as Lagrange's equations (see below).

Whatever the method used to derive the equation of motions, the dynamical equations of motion for multibody systems in terms of generalized coordinates $\{q_k(t)\}$ have the form

$$\sum m_{ij}\ddot{q}_j + \sum\sum \mu_{ijk}\dot{q}_j\dot{q}_k = Q_i \tag{7.10}$$

The first term on the left involves a generalized symmetric mass matrix $m_{ij} = m_{ji}$. The second term includes Coriolis and centripetal acceleration. The right-hand side includes all the force and control terms. This equation has a quadratic nonlinearity in the generalized velocities. These quadratic terms usually drop out for rigid body problems with a single axis of rotation. However, the nonlinear inertia terms generally appear in problems with simultaneous rotation about two or three axes as in multi-link robot arms (Fig. 7.2), gyroscope problems, and slewing momentum wheels in satellites.

In modern dynamic simulation software, called multibody codes, these equations are automatically derived and integrated once the user specifies the geometry, forces, and controls. Some of these codes are called ADAMS, DADS, Working Model, and NEWEUL. However, the designer must use caution as these codes are sometimes poor at modeling friction and impacts between bodies.

7.5 Simple Dynamic Models

Two simple examples of the application of the angular momentum law are now given. The first is for rigid body rotation about a single axis and the second has two axes of rotation.

Compound Pendulum

When a body is constrained to a single rotary degree of freedom and is acted on by the force of gravity as in Fig. 7.5, the equation of motion takes the form, where θ is the angle from the vertical,

$$I\vartheta - (m_1L_1 - m_2L_2)g \sin\theta = T(t) \tag{7.11}$$

where $T(t)$ is the applied torque, $I = m_1L_1^2 + m_2L_2^2$ is the moment of inertia (properly called the second moment of mass). The above equation is nonlinear in the sine function of the angle. In the case of small motions about $\theta = 0$, the equation becomes a linear differential equation and one can look for solutions of the form $\theta = A\cos\omega t$, when $T(t) = 0$. For this case the pendulum exhibits sinusoidal motion with

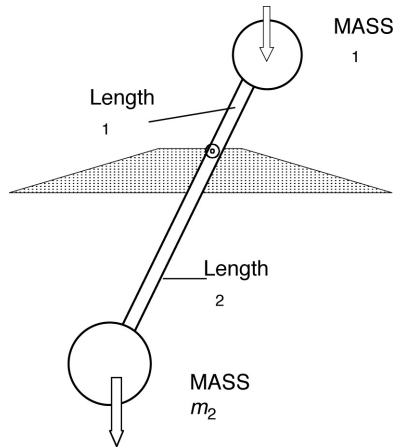



FIGURE 7.5 Sketch of a compound pendulum under gravity torques.

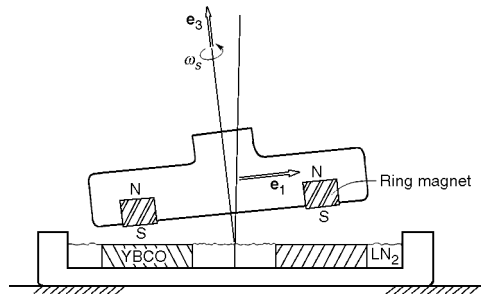



FIGURE 7.6 Sketch of a magnetically levitated flywheel on high-temperature superconducting bearings.

natural frequency

$$\omega = [g(m_2L_2 - m_1L_1)/I]^{1/2} \tag{7.12}$$

For the simple pendulum $m_1 = 0$, and we have the classic pendulum relation in which the natural frequency depends inversely on the square root of the length:

$$\omega = (g/L_2)^{1/2} \tag{7.13}$$

Gyroscopic Motions

Spinning devices such as high speed motors in robot arms or turbines in aircraft engines or magnetically levitated flywheels (Fig. 7.6) carry angular momentum, devoted by the vector **H**. Euler's extension of Newton's laws says that a change in angular momentum must be accompanied by a force moment **M**,

$$\mathbf{M} = \dot{\mathbf{H}} \tag{7.14}$$

In three-dimensional problems one can often have components of angular momentum about two different axes. This leads to a Coriolis acceleration that produces a gyroscopic moment even when the two angular motions are steady. Consider the spinning motor with spin ϕ about an axis with unit vector $\mathbf{e}_1$ and

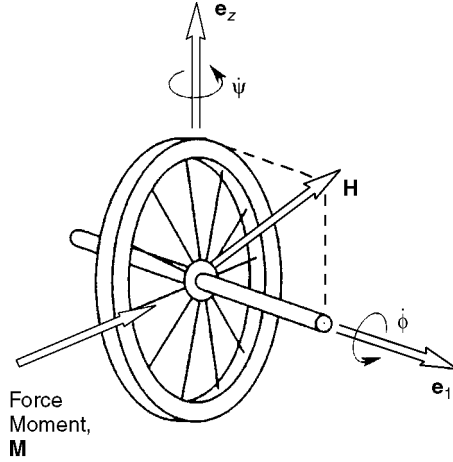



FIGURE 7.7 Gyroscopic moment on a precessing, spinning rigid body.

let us imagine an angular motion of the $\mathbf{e}_1$ axis, ψ about a perpendicular axis $\mathbf{e}_z$ called the precession axis in gyroscope parlance. Then one can show that the angular momentum is given by

$$\mathbf{H} = I_1\phi\mathbf{e}_1 + I_z\psi\mathbf{e}_z \tag{7.15}$$

and the rate of change of angular momentum for constant spin and presession rates is given by

$$\dot{\mathbf{H}} = \psi\mathbf{e}_z \times \mathbf{H} \tag{7.16}$$

There must then exist a gyroscopic moment, often produced by forces on the bearings of the axel (Fig. 7.7). This moment is perpendicular to the plane formed by $\mathbf{e}_1$ and $\mathbf{e}_z$, and is proportional to the product of the rotation rates:

$$\mathbf{M} = I_1\phi\psi\mathbf{e}_z \times \mathbf{e}_1 \tag{7.17}$$

This has the same form as Eq. (7.10), when the generalized force Q is identified with the moment **M**, i.e., the moment is the product of generalized velocities when the second derivative acceleration terms are zero.

7.6 Elastic System Modeling

Elastic structures take the form of cables, beams, plates, shells, and frames. For linear problems one can use the method of eigenmodes to represent the dynamics with a finite set of modal amplitudes for generalized degrees of freedom. These eigenmodes are found as solutions to the PDEs of the elastic structure (see, e.g., Yu, 1996).

The simplest elastic structure after the cable is a one-dimensional beam shown in Fig. 7.8. For small motions we assume only transverse displacements $w(x, t)$, where x is a spatial coordinate along the beam. One usually assumes that the stresses on the beam cross section can be integrated to obtain stress vector resultants of shear V, bending moment M, and axial load T. The beam can be loaded with point or concentrated forces, end forces or moment or distributed forces as in the case of gravity, fluid forces, or electromagnetic forces. For a distributed transverse load $f(x, t)$, the equation of motion is given by

$$D\frac{\partial^4 w}{\partial x^4} - T\frac{\partial^2 w}{\partial x^2} + \rho A\frac{\partial^2 w}{\partial t^2} = f(x, t) \tag{7.18}$$

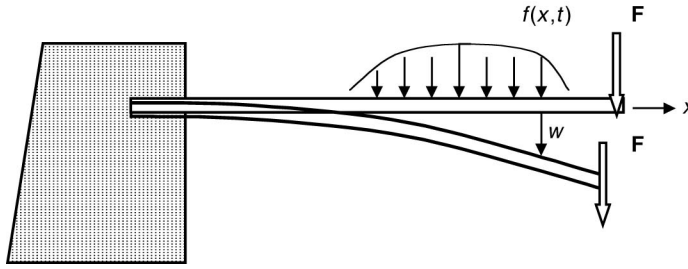



FIGURE 7.8 Sketch of an elastic cantilevered beam.

where D is the bending stiffness, A is the cross-sectional area of the beam, and ρ is the density. For a beam with Young's modulus Y, rectangular cross section of width b, and height h, $D = Ybh^3/12$. For $D = 0$, one has a cable or string under tension T, and the equation takes the form of the usual wave equation. For a beam with tension T, the natural frequencies are increased by the addition of the second term in the equation. For $T = -P$, i.e., a compressive load on the end of the beam, the curvature term leads to a decrease of natural frequency with increase of the compressive force P. If the lowest natural frequency goes to zero with increasing load P, the straight configuration of the beam becomes unstable or undergoes *buckling*. The use of T or $(-P)$ to stiffen or destiffen a beam structure can be used in design of sensors to create a sensor with variable resonance. This idea has been used in a MEMS accelerometer design (see below).

Another feature of the beam structure dynamics is the fact that unlike the string or cable, the frequencies of the natural modes are not commensurate due to the presence of the fourth-order derivative term in the equation. In wave type problems this is known as *wave dispersion*. This means that waves of different wavelengths travel at different speeds so that wave pulse shapes change their form as the wave moves through the structure.

In order to solve dynamic problems in finite length beam structures, one must specify boundary conditions at the ends. Examples of boundary conditions include

$$\begin{aligned}
&\text{clamped end} && w = 0, \quad \frac{\partial w}{\partial x} = 0 \\
&\text{pinned end} && w = 0, \quad \frac{\partial^2 w}{\partial x^2} = 0 \ \text{(zero moment)} \\
&\text{free end} && \frac{\partial^2 w}{\partial x^2} = 0, \quad \frac{\partial^3 w}{\partial x^3} = 0 \ \text{(zero shear)}
\end{aligned} \tag{7.19}$$

Piezoelastic Beam

Piezoelastic materials exhibit a coupling between strain and electric polarization or voltage. Thus, these materials can be used for sensors or actuators. They have been used for active vibration suppression in elastic structures. They have also been explored for active optics space applications. Many natural materials exhibit piezoelasticity such as quartz as well as manufactured materials such as barium titanate, lead zirconate titanate (PZT), and polyvinylidene fluoride (PVDF). Unlike forces on charges and currents (see below), the electric effect takes place through a change in shape of the material. The modeling of these devices can be done by modifying the equations for elastic structures.

The following work on piezo-benders is based on the work of Lee and Moon (1989) as summarized in Miu (1993). One of the popular configurations of a piezo actuator-sensor is the piezo-bender shown in Fig. 7.9. The elastic beam is of rectangular cross section as is the piezo element. The piezo element

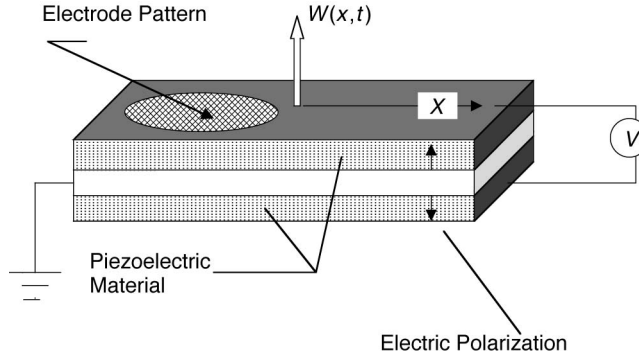



FIGURE 7.9 Elastic beam with two piezoelectric layers (Lee and Moon, 1989).

can be cemented on one or both sides of the beam either partially or totally covering the surface of the non-piezo substructure.

In general the local electric dipole polarization depends on the six independent strain components produced by normal and shear stresses. However, we will assume that the transverse voltage or polarization is coupled to the axial strain in the plate-shaped piezo layers. The constitutive relations between axial stress and strain, T, S, electric field and electric displacement, E_3, D_3 (not to be confused with the bending stiffness D), are given by

$$T_1 = c_{11}S_1 - e_{31}E_3, \qquad D_3 = e_{31}S_1 + \varepsilon_3 E_3 \tag{7.20}$$

The constants c_{11}, e_{31}, ε_3, are the elastic stiffness modulus, piezoelectric coupling constant, and the electric permittivity, respectively.

If the piezo layers are polled in the opposite directions, as shown in the Fig. 7.9, an applied voltage will produce a strain extention in one layer and a strain contraction in the other layer, which has the effect of an applied moment on the beam. The electrodes applied to the top and bottom layers of the piezo layers can also be shaped so that there can be a gradient in the average voltage across the beam width. For this case the equation of motion of the composite beam can be written in the form

$$D\frac{\partial^4 w}{\partial x^4} + \rho A\frac{\partial^2 w}{\partial t^2} = -2e_{31}z_o\frac{\partial^2 V_3}{\partial x^2} \tag{7.21}$$

where $z_o = (h_S + h_P)/2$.

The z term is the average of piezo plate and substructure thicknesses. When the voltage is uniform, then the right-hand term results in an applied moment at the end of the beam proportional to the transverse voltage.

7.7 Electromagnetic Forces

One of the keys to modeling mechatronic systems is the identification of the electric and magnetic forces. Electric forces act on charges and electric polarization (electric dipoles). Magnetic forces act on electric currents and magnetic polarization. Electric charge and current can experience a force in a uniform electric or magnetic field; however, electric and magnetic dipoles will only produce a force in an electric or magnetic field gradient.

Electric and magnetic forces can also be calculated using both direct vector methods as well as from energy principles. One of the more popular methods is *Lagrange's equation* for electromechanical systems described below.

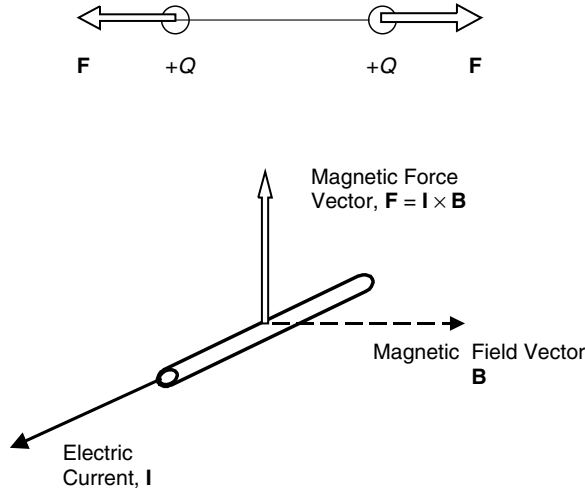



FIGURE 7.10 Electric forces on two charges (top). Magnetic force on a current carrying wire element (bottom).

Electromagnetic systems can be modeled as either distributed field quantities, such as electric field **E** or magnetic flux density **B** or as lumped element electric and magnetic circuits. The force on a point charge Q is given by the vector equation (Fig. 7.10):

$$\mathbf{F} = Q\mathbf{E} \tag{7.22}$$

When **E** is generated by a single charge, the force between charges Q_1 and Q_2 is given by

$$F = \frac{Q_1 Q_2}{4\pi\varepsilon_0 r^2} \tag{7.23}$$

and is directed along the line connecting the two charges. Like charges repel and opposite charges attract one another.

The magnetic force per unit length on a current element **I** is given by the cross product

$$\mathbf{F} = \mathbf{I} \times \mathbf{B} \tag{7.24}$$

where the magnetic force is perpendicular to the plane of the current element and the magnetic field vector. The total force on a closed circuit in a uniform field can be shown to be zero. Net forces on closed circuits are produced by field gradients due to other current circuits or field sources.

Forces produced by field distributions around a volume containing electric charge or current can be calculated using the field quantities of **E**, **B** directly using the concept of magnetic and electric stresses, which was developed by Faraday and Maxwell. These electromagnetic stresses must be integrated over an area surrounding the charge or current distribution. For example, a solid containing a current distribution can experience a *magnetic pressure*, $P = B_t^2/2\mu_0$, on the surface element and a *magnetic tension*, $t_n = B_n^2/2\mu_0$, where the magnetic field components are written in terms of values tangential and normal to the surface. Thus, a one-tesla magnetic field outside of a solid will experience 40 N/cm^2 pressure if the field is tangential to the surface.

In general there are four principal methods to calculate electric and magnetic forces:

- direct force vectors and moments between electric charges, currents, and dipoles;
- electric field-charge and magnetic field-current force vectors;

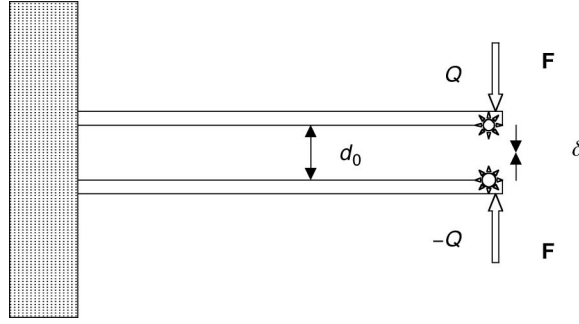



FIGURE 7.11 Two elastic beams with electric charges at the ends.

- electromagnetic tensor, integration of electric tension, magnetic pressure over the surface of a material body; and
- energy methods based on gradients of magnetic and electric energy.

Examples of the direct method and stress tensor method are given below. The energy method is described in the section on Lagrange's equations.

Example 1. Charge–Charge Forces

Suppose two elastic beams in a MEMS device have electric charges Q_1, Q_2 coulombs each concentrated at their tips (Fig. 7.11). The electric force between the charges is given by the vector

$$\mathbf{F} = \frac{Q_1 Q_2}{4\pi\varepsilon_0}\frac{\mathbf{r}}{r^3} \quad \text{(newtons)} \tag{7.25}$$

where $1/4\pi\varepsilon_0 = 8.99 \times 10^9\ \text{Nm}^2/\text{C}^2$.

If the initial separation between the beams is d_0, we seek the new separation under the electric force. For simplicity, we let $Q_1 = -Q_2 = Q$, where opposite charges create an attractive force between the beam tips. The deflection of the cantilevers is given by

$$\delta = \frac{FL^3}{3YI} = \frac{1}{k}F \tag{7.26}$$

where L is the length, Y the Young's modulus, I the second moment of area, and k the effective spring constant.

Under the electric force, the new separation is $d = d_0 - 2\delta$,

$$k\delta = \frac{Q^2}{4\pi\varepsilon_0}\frac{1}{(d_0 - 2\delta)^2} \tag{7.27}$$

For $\delta \ll d_0$ to first order we have

$$\delta = \frac{Q^2/4\pi\varepsilon_0 d_0^2 k}{1 - (1/d_0^3)(Q^2/k\pi\varepsilon_0)} \tag{7.28}$$

This problem shows the potential for electric field buckling because as the beam tips move closer together, the attractive force between them increases. The nondimensional expression in the denominator

$$\frac{Q^2}{\pi\varepsilon_0 d_0^3}\frac{1}{k} \tag{7.29}$$

is the ratio of the negative electric stiffness to the elastic stiffness k of the beams.

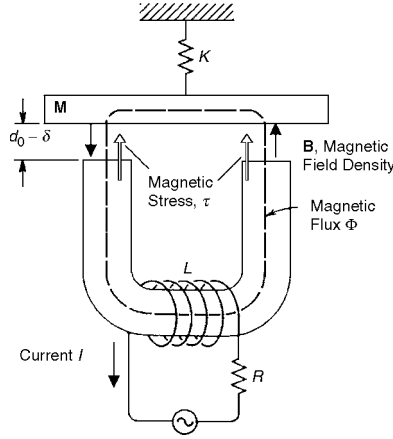



FIGURE 7.12 Force on a ferromagnetic bar near an electromagnet.

Example 2. Magnetic Force on an Electromagnet

Imagine a ferromagnetic keeper on an elastic restraint of stiffness *k,* as shown in Fig. 7.12. Under the soft magnetic keeper, we place an electromagnet which produces N turns of current I around a soft ferromagnetic core. The current is produced by a voltage in a circuit with resistance R.

The magnetic force will be calculated using the *magnetic stress tensor* developed by Maxwell and Faraday (see, e.g., Moon, 1984, 1994). Outside a ferromagnetic body, the stress tensor is given by **t** and the stress vector on the surface defined by normal **n** is given by $\boldsymbol{\tau} = \mathbf{t} \cdot \mathbf{n}$:

$$\tau = \frac{1}{\mu_0}\left(\frac{1}{2}[B_n^2 - B_t^2],\ B_n B_t\right) = (\tau_n, \tau_t) \tag{7.30}$$

For high magnetic permeability as in a ferromagnetic body, the tangential component of the magnetic field outside the surface is near zero. Thus the force is approximately normal to the surface and is found from the integral of the magnetic tension over the surface:

$$\mathbf{F} = \frac{1}{2\mu_0}\int B_n^2 \mathbf{n}\ dA \tag{7.31}$$

and $B_n^2/2\mu_0$ represents a magnetic tensile stress. Thus, if the area of the pole pieces of the electromagnet is A (neglecting fringing of the field), the force is

$$F = B_g^2 A/\mu_0 \tag{7.32}$$

where B_g is the gap field. The gap field is determined from Amperes law

$$NI = \widehat{R}\,\Phi, \quad \Phi = B_g A \tag{7.33}$$

where the *reluctance* is approximately given by

$$\widehat{R} = \frac{2(d_0 - \delta)}{\mu_0 A} \tag{7.34}$$

The balance of magnetic and elastic forces is then given by

$$F = \frac{1}{\mu_0 A}\Phi^2 = \frac{1}{\mu_0 A}\left(\frac{NI}{\widehat{R}}\right)^2 = k\delta \tag{7.35}$$

or

$$\frac{(NI)^2}{4(d_0 - \delta)^2}\mu_0 A = k\delta, \qquad \frac{\mu_0 N^2 I^2 A}{4(d_0 - \delta)^2} = k\delta$$

(Note that the expression $\mu_0 N^2 I^2$ has units of force.) Again as the current is increased, the total elastic and electric stiffness goes to zero and one has the potential for buckling.

7.8 Dynamic Principles for Electric and Magnetic Circuits

The fundamental equations of electromagnetics stem from the work of nineteenth century scientists such as Faraday, Henry, and Maxwell. They take the form of partial differential equations in terms of the field quantities of electric field **E** and magnetic flux density **B**, and also involve volumetric measures of charge density q and current density **J** (see, e.g., Jackson, 1968). Most practical devices, however, can be modeled with lumped electric and magnetic circuits. The standard resistor, capacitor, inductor circuit shown in Fig. 7.13 uses electric current I (amperes), charge Q (columbs), magnetic flux Φ (webers), and voltage V (volts) as dynamic variables. The voltage is the integral of the electric field along a path:

$$V_{21} = \int_1^2 \mathbf{E} \cdot dl \tag{7.36}$$

The charge Q is the integral of charge density q over a volume, and electric current I is the integral of normal component of **J** across an area. The magnetic flux Φ is given as another surface integral of magnetic flux.

$$\Phi = \int \mathbf{B} \cdot d\mathbf{A} \tag{7.37}$$

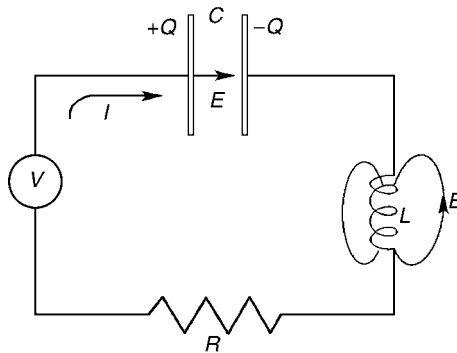



FIGURE 7.13 Electric circuit with lumped parameter capacitance, inductance, and resistance.

When there are no mechanical elements in the system, the dynamical equations take the form of conservation of charge and the Faraday–Henry law of flux change.

$$\frac{dQ}{dt} = I \quad \text{(Conservation of charge)} \tag{7.38}$$

$$\frac{d\phi}{dt} = V \quad \text{(Law of flux change)} \tag{7.39}$$

where $\phi = N\Phi$ is called the number of flux linkages, and N is an integer. In electromagnetic circuits the analog of mechanical constitutive properties is inductance L and capacitance C. The magnetic flux in an inductor, for example, often depends on the current I.

$$\phi = f(I) \tag{7.40}$$

For a linear inductor we have a definition of inductance L, i.e., $\phi = LI$. If the system has a mechanical state variable such as displacement x, as in a magnetic solenoid actuator, then L may be a function of x.

In charge storage circuit elements, the capacitance C is defined as

$$Q = CV \tag{7.41}$$

In MEMS devices and in microphones, the capacitance may also be a function of some generalized mechanical displacement variable.

The voltages across the different circuit elements can be active or passive. A pure voltage source can maintain a given voltage, but the current depends on the passive voltages across the different circuit elements as summarized in the Kirchhoff circuit law:

$$\frac{d}{dt}L(x)I + \frac{Q}{C(x)} + RI = V(t) \tag{7.42}$$

Lagrange's Equations of Motion for Electromechanical Systems

It is well known that the Newton–Euler equations of motion for mechanical systems can be derived using an energy principle called Lagrange's equation. In this method one identifies generalized coordinates $\{q_k\}$, not to be confused with electric charges, and writes the kinetic energy of the system T in terms of generalized velocities and coordinates, $T(\dot{q}_k, q_k)$. Next the mechanical forces are split into so-called conservative forces, which can be derived from a potential energy function $W(q_k)$ and the rest of the forces, which are represented by a generalized force Q_k corresponding to the work done by the kth generalized coordinate. Lagrange's equations for mechanical systems then take the form:

$$\frac{d}{dt}\frac{\partial T(\dot{q}_k, q_k)}{\partial \dot{q}_k} - \frac{\partial T}{\partial q_k} + \frac{\partial W(q_k)}{\partial q_k} = Q_k \tag{7.43}$$

For example, in a linear spring–mass–damper system, with mass m, spring constant k, viscous damping constant c, and one generalized coordinate $q_1 = x$, the equation of motion can be derived using, $T = \frac{1}{2}m\dot{x}^2$, $W = \frac{1}{2}kx^2$, $Q_1 = -c\dot{x}$, in Lagrange's equation above. What is remarkable about this formulation is that it can be extended to treat both electromagnetic circuits and coupled electromechanical problems.

As an example of the application of Lagrange's equations to a coupled electromechanical problem, consider the one-dimensional mechanical device, shown in Fig. 7.14, with a magnetic actuator and a capacitance actuator driven by a circuit with applied voltage $V(t)$. We can extend Lagrange's equation to

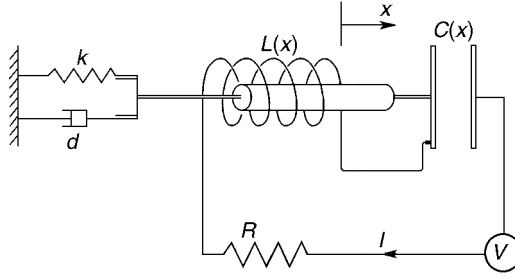



FIGURE 7.14 Coupled lumped parameter electromechanical system with single degree of freedom mechanical motion $x(t)$.

circuits by defining the charge on the capacitor, Q, as another generalized coordinate along with x, i.e., in Lagrange's formulation, $q_1 = x$, $q_2 = Q$. Then we add to the kinetic energy function a magnetic energy function $W_m(\dot{Q}, x)$, and add to the potential energy an electric field energy function $W_e(Q, x)$. The equations of both the mass and the circuit can then be derived from

$$\frac{d}{dt}\frac{\partial[T + W_m]}{\partial \dot{q}_k} - \frac{\partial[T + W_m]}{\partial q_k} + \frac{\partial[W + W_e]}{\partial q_k} = Q_k \tag{7.44}$$

The generalized force must also be modified to account for the energy dissipation in the resistor and the energy input of the applied voltage $V(t)$, i.e., $Q_1 = -c\dot{x}$, $Q_2 = -R\dot{Q} + V(t)$. In this example the magnetic energy is proportional to the inductance $L(x)$, and the electric energy function is inversely proportional to the capacitance $C(x)$. Applying Lagrange's equations automatically results in expressions for the magnetic and electric forces as derivatives of the magnetic and electric energy functions, respectively, i.e.,

$$W_m = \frac{1}{2}L(x)\dot{Q}^2 = \frac{1}{2}LI^2, \qquad W_e = \frac{1}{2C(x)}Q^2 \tag{7.45}$$

$$F_m = \frac{\partial W_m(x, \dot{Q})}{\partial x} = \frac{1}{2}I^2\frac{dL(x)}{dx}, \qquad F_e = -\frac{\partial W_e(x, Q)}{\partial x} = -\frac{1}{2}Q^2\frac{d}{dx}\left[\frac{1}{C(x)}\right] \tag{7.46}$$

These remarkable formulii are very useful in that one can calculate the electromagnetic forces by just knowing the dependence of the inductance and capacitance on the displacement x. These functions can often be found from electrical measurements of L and C.

Example: Electric Force on a Comb-Drive MEMS Actuator

Consider the motion of an elastically constrained plate between two grounded fixed plates as in a MEMS comb-drive actuator in Fig. 7.15. When the moveable plate has a voltage V applied, there is stored electric field energy in the two gaps given by

$$W_e^*(V, x) = \frac{1}{2}\varepsilon_0 V^2 A \frac{d_0}{d_0^2 - x^2} \tag{7.47}$$

In this expression the electric energy function is written in terms of the voltage V instead of the charge on the plates Q as in Eqs. (7.45) and (7.46). Also the initial gap is d_0, and the area of the plate is A.

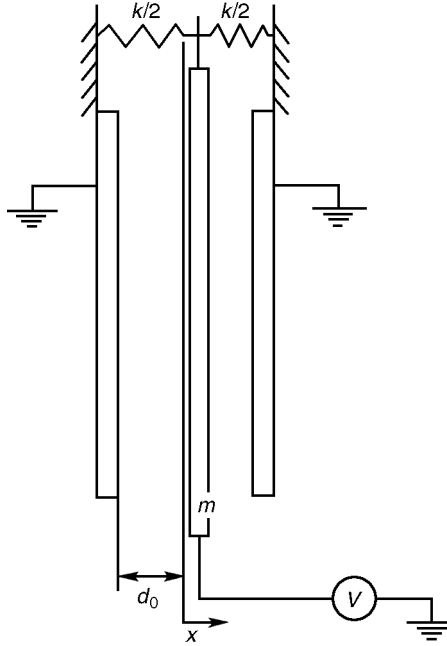



FIGURE 7.15 Example of electric force on the elements of a comb-drive actuator.

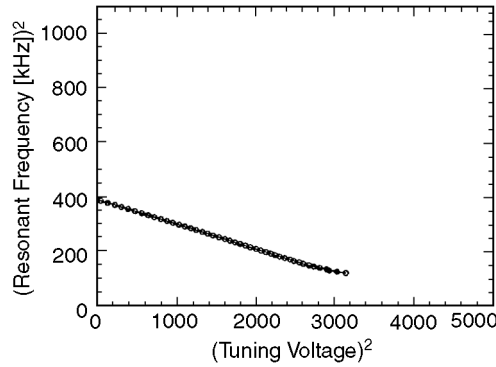



FIGURE 7.16 Decrease in natural frequency of a MEMS device with applied voltage as an example of negative electric stiffness [From Adams (1996)].

Using the force expressions derived from Lagrange's equations (7.44), the electric charge force on the plate is given by

$$F_e = \frac{\partial}{\partial x} W_e^*(V, x) = \frac{\varepsilon_0 V^2 A}{d_0} \frac{x}{(1 - x^2/d^2)^2} \tag{7.48}$$

This expression shows that the electric stiffness is negative for small x, which means that the voltage will decrease the natural frequency of the plate. This idea has been applied to a MEMS comb-drive actuator by Adams (1996) in which the voltage could be used to tune the natural frequency of a MEMS accelerometer, as shown in Fig. 7.16.

7.9 Earnshaw's Theorem and Electromechanical Stability

It is not well known that electric and magnetic forces in mechanical systems can produce static instability, otherwise known as *elastic buckling* or *divergence*. This is a consequence of the inverse square nature of many electric and magnetic forces. It is well known that the electric and magnetic field potential Φ satisfies Laplace's equation, $\nabla^2\Phi = 0$. There is a basic theorem in potential theory about the impossibility of a relative maximum or minimum value of a potential $\Phi(\mathbf{r})$ for solutions of Laplace's equation except at a boundary. It was stated in a theorem by Earnshaw (1829) that it is impossible for a static set of charges, magnetic and electric dipoles, and steady currents to be in a stable state of equilibrium without mechanical or other feedback or dynamic forces (see, for example, Moon, 1984, 1994).

One example of Earnshaw's theorem is the instability of a magnetic dipole (e.g., a permanent magnet) near a ferromagnetic surface (Fig. 7.17). Levitated bearings based on ferromagnetic forces, for example, require feedback control. Earnshaw's theorem also implies that if there is one degree of freedom with stable restoring forces, there must be another degree of freedom that is unstable. Thus the equilibrium positions for a pure electric or magnetic system of charges and dipoles must be saddle points. The implication for the force potentials is that the matrix of second derivatives is not positive definite. For example, suppose there are three generalized position coordinates $\{s_u\}$ for a set of electric charges. Then if the generalized forces are proportional to the gradient of the potential, $\nabla\Phi$, then the generalized electric stiffness matrix $\mathbf{K}_{ij}$, given by

$$\mathbf{K}_{ij} = \left[\frac{\partial^2\Phi}{\partial s_i \partial s_j}\right]$$

will not be positive definite. This means that at least one of the eigenvalues will have negative stiffness.

Another example of electric buckling is a beam in an electric field with charge induced by an electric field on two nearby stationary plates as in Fig. 7.15. The induced charge on the beam will be attracted to either of the two plates, but is resisted by the elastic stiffness of the beam. As the voltage is increased, the combined electric and elastic stiffnesses will decrease until the beam buckles to one or the other of the two sides. Before buckling, however, the natural frequency of the charged beam will decrease (Fig. 7.16). This property has been observed experimentally in a MEMS device. A similar magneto elastic buckling is observed for a thin ferromagnetic elastic beam in a static magnetic field (see Moon, 1984). Both electroelastic and magnetoelastic buckling are derived from the same principle of Earnshaw's theorem.

There are dramatic exceptions to Earnshaw's stability theorem. One of course is the levitation of 50-ton vehicles with magnetic fields, known as MagLev, or the suspension of gas pipeline rotors using feedback controlled magnetic bearings (see Moon, 1994). Here either the device uses feedback forces, i.e., the fields

Ferromagnetic Material

S

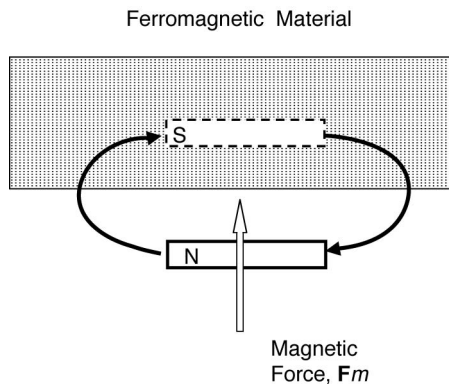
N

Magnetic Force, **F***m*

FIGURE 7.17 Magnetic force on a magnetic dipole magnet near a ferromagnetic half space with image dipole shown.

are not static, or the source of one of the magnetic fields is a superconductor. Diamagnetic forces are exceptions to Earnshaw's theorem, and superconducting materials have properties that behave like diamagnetic materials. Also new high-temperature superconductivity materials, such as YBaCuO, exhibit magnetic flux pinning forces that can be utilized for stable levitation in magnetic bearings without feedback (see Moon, 1994).

References

Adams, S. G. (1996), *Design of Electrostatic Actuators to Tune the Effective Stiffness of Micro-Mechanical Systems,* Ph.D. Dissertation, Cornell Unversity, Ithaca, New York.

Goldstein, H. (1980), *Classical Mechanics,* Addison-Wesley, Reading, MA.

Jackson, J. D. (1968), *Classical Electrodynamics,* J. Wiley & Sons, New York.

Lee, C. K. and Moon, F. C. (1989), "Laminated piezopolymer plates for bending sensors and actuators," *J. Acoust. Soc. Am.*, **85**(6), June 1989.

Melcher, J. R. (1981), *Continuum Electrodynamics,* MIT Press, Cambridge, MA.

Miu, D. K. (1993), *Mechatronics,* Springer-Verlag, New York.

Moon, F. C. (1984), *Magneto-Solid Mechanics,* J. Wiley & Sons, New York.

Moon, F. C. (1994), *Superconducting Levitation,* J. Wiley & Sons, New York.

Moon, F. C. (1999), *Applied Dynamics,* J. Wiley & Sons, New York.

Yu, Y.-Y. (1996), *Vibrations of Elastic Plates,* Springer-Verlag, New York.

8

Structures and Materials

Eniko T. Enikov
University of Arizona

The term mechatronics was first used by Japanese engineers to define a mechanical system with embedded electronics, capable of providing intelligence and control functions. Since then, the continued progress in integration has led to the development of microelectromechanical systems (MEMS) in which the mechanical structures themselves are part of the electrical subsystem. The development and design of such mechatronic systems requires interdisciplinary knowledge in several disciplines—electronics, mechanics, materials, and chemistry. This section contains an overview of the main mechanical structures, the materials they are built from, and the governing laws describing the interaction between electrical and mechanical processes. It is intended for use in the initial stage of the design, when quick estimates are necessary to validate or reject a particular concept. Special attention is devoted to the newly emerging smart materials—electroactive polymer actuators. Several tables of material constants are also provided for reference.

8.1 Fundamental Laws of Mechanics

Statics and Dynamics of Mechatronic Systems

The fundamental laws of mechanics are the balance of linear and angular momentum. For an idealized system consisting of a point mass m moving with velocity $\mathbf{v}$, the linear momentum is defined as the product of the mass and the velocity:

$$\mathbf{L} = m\mathbf{v} \tag{8.1}$$

The conservation of linear momentum for a single particle postulates that the rate of change of linear momentum is equal to the sum of all forces acting on the particle

$$\dot{\mathbf{L}} = m\dot{\mathbf{v}} = \sum \mathbf{F}_i \tag{8.2}$$

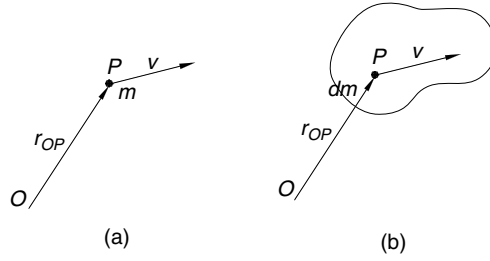



FIGURE 8.1 Difinition of velocity and position vectors for single particle (a) and rigid body (b).

where we have assumed that the mass does not change over time. The angular momentum of a particle with respect to an arbitrary reference point O is defined as

$$\mathbf{H}_O = \mathbf{r}_{OP} \times (m\mathbf{v}) \tag{8.3}$$

where $\mathbf{r}_{OP}$ is the position vector between points O and P (see Fig. 8.1(a)). The balance of angular momentum for a single infinitesimally small particle is automatically satisfied as a result of (8.1). In the case of multiple particles (a rigid body composed of infinite number of particles), the linear and angular momenta are defined as the sum (integral) of the momentum of individual particles (Fig. 8.1(b)):

$$\mathbf{L} = \int_V \mathbf{v}\, dm \qquad \text{and} \qquad \mathbf{H}_O = \int_V \mathbf{r}_{OP} \times \mathbf{v}\, dm \tag{8.4}$$

The second fundamental law of classical mechanics states that the rate of change of angular momentum is equal to the sum of all moments acting on the body:

$$\dot{\mathbf{H}}_O = \sum_i \mathbf{M}_i + \sum_i \mathbf{r}_i \times \mathbf{F}_i \tag{8.5}$$

where $\mathbf{M}_i$ are the applied external force-couples in addition to the forces $\mathbf{F}_i$. If the reference point O is chosen to be the center of mass of the body G, the linear and angular momentum balance law take a simpler form:

$$m\dot{\mathbf{v}}_G = \sum_i \mathbf{F}_i \tag{8.6}$$

$$I_G\dot{\boldsymbol{\omega}} = \sum_j \mathbf{M}_j + \sum_i \mathbf{r}_i \times \mathbf{F}_i \tag{8.7}$$

where $\boldsymbol{\omega}$ is the instantaneous vector of angular velocity and I_G is the moment of inertia about the center of mass. Equations (8.6) and (8.7) are called equations of motion and play a central role in the dynamics of rigid bodies. If there is no motion (linear and angular velocities are zero), one is faced with a *statics problem*. Conversely, when the accelerations are large, we need to solve the complete system of Eqs. (8.6) and (8.7) including the inertial terms. In mechatronic systems the mechanical response is generally slower than the electrical one and therefore determines the overall response. If the response time is critical to the application, one needs to consider the inertial terms in Eqs. (8.6) and (8.7).

Equations of Motion of Deformable Bodies

Rigid bodies do not change shape or size during their motion, that is, the distance between the particles they are made of is constant. In reality, all objects deform to a certain extent when subjected to external forces. Whether a body can be treated as rigid or deformable is dictated by the particular application.

In this section we will review the fundamental equations describing the motion of deformable bodies. These equations also result from the balance of linear and angular momentum applied to an infinitesimally small portion of the material volume dV. Each element dV is subjected not only to external body force $\mathbf{f}$, but also to internal forces originating from the rest of the body. These internal forces are described by a second order tensor $\mathbf{T}$, called stress tensor. The balance of linear momentum can then be stated in integral form for an *arbitrary* portion of the body occupying volume V as

$$\frac{d}{dt}\int_V \rho\mathbf{v}\,dv = \int_{\partial V}\mathbf{T}\cdot\mathbf{n}\,dA + \int_V \mathbf{f}\,dv \tag{8.8}$$

where ρ is the mass density, $\mathbf{v}$ is the velocity of the element dV, and $\mathbf{f}$ is the force per unit volume acting upon dV. The above balance law states that the rate of change of linear momentum is equal to the sum of the internal force flux (stress) acting on the boundary of V and the external body force, distributed inside V. Applying the transport theorem to (8.8) along with the mass conservation law reduces the above to

$$\int_V \rho\dot{\mathbf{v}}\,dv = \int_V \nabla\cdot\mathbf{T}\,dv + \int_V \mathbf{f}\,dv \tag{8.9}$$

Since (8.9) is valid for an arbitrary volume, it follows that the integrands are also equal. Thus the local (differential) form of linear momentum balance is

$$\rho\dot{\mathbf{v}} = \nabla\cdot\mathbf{T} + \mathbf{f} \qquad \text{or with index notation} \qquad \rho\dot{v}_i = T_{ij,j} + f_i \tag{8.10}$$

Using analogous procedure, the balance of angular momentum can be shown to reduce to a simple symmetry condition of the stress tensor

$$T_{ij} = T_{ji} \tag{8.11}$$

which is valid for materials without external body couples. It should be mentioned that in certain anisotropic materials, the polarization or magnetization vectors can develop body couples, for example when $\mathbf{E}\times\mathbf{P}\neq\mathbf{0}$. In these cases the stress tensor is nonsymmetric and its vector invariant is equal to the body couple. Equations (8.10) are usually used in one of the three most common coordinate systems. For example, using rectangular coordinates we have

$$\begin{aligned}
&\frac{\partial T_{xx}}{\partial x} + \frac{\partial T_{xy}}{\partial y} + \frac{\partial T_{xz}}{\partial z} + f_x = \rho a_x, \quad T_{xy} = T_{yx}\\
&\frac{\partial T_{yx}}{\partial x} + \frac{\partial T_{yy}}{\partial y} + \frac{\partial T_{yz}}{\partial z} + f_y = \rho a_y, \quad T_{yz} = T_{zy}\\
&\frac{\partial T_{zx}}{\partial x} + \frac{\partial T_{zy}}{\partial y} + \frac{\partial T_{zz}}{\partial z} + f_z = \rho a_z, \quad T_{xz} = T_{zx}
\end{aligned} \tag{8.12}$$

and in cylindrical coordinates

$$\begin{aligned}
&\frac{\partial T_{rr}}{\partial r} + \frac{T_{rr}-T_{\theta\theta}}{r} + \frac{1}{r}\frac{\partial T_{r\theta}}{\partial\theta} + \frac{\partial T_{rz}}{\partial z} + f_r = \rho a_r, \quad T_{r\theta} = T_{\theta r}\\
&\frac{\partial T_{r\theta}}{\partial r} + \frac{2}{r}T_{r\theta} + \frac{1}{r}\frac{\partial T_{\theta\theta}}{\partial\theta} + \frac{\partial T_{\theta z}}{\partial z} + f_\theta = \rho a_\theta, \quad T_{\theta z} = T_{z\theta}\\
&\frac{\partial T_{rz}}{\partial r} + \frac{1}{r}T_{rz} + \frac{1}{r}\frac{\partial T_{\theta z}}{\partial\theta} + \frac{\partial T_{zz}}{\partial z} + f_z = \rho a_z, \quad T_{rz} = T_{zr}
\end{aligned} \tag{8.13}$$

where (x, y, z) and (r, θ, z) are the three coordinates, f's are the corresponding body force densities, and a's are the accelerations. In addition to Eqs. (8.12) or (8.13), a relation between the stress and the displacement is needed in order to determine the deformation. Since the rigid body translations and

rotations do not cause deformation of the body, they do not affect the internal stress field either. In fact, the latter is a function of the gradient of the displacement, called deformation gradient. When this gradient is small, a linear relationship between the displacements and strains can be used

$$\varepsilon_x = \frac{\partial u_x}{\partial x}, \quad \varepsilon_y = \frac{\partial u_y}{\partial y}, \quad \varepsilon_z = \frac{\partial u_z}{\partial z}, \quad \varepsilon_{xy} = \frac{\partial u_x}{\partial y} + \frac{\partial u_y}{\partial x}, \quad \varepsilon_{xz} = \frac{\partial u_x}{\partial z} + \frac{\partial u_z}{\partial x}, \quad \varepsilon_{zy} = \frac{\partial u_z}{\partial y} + \frac{\partial u_y}{\partial z} \quad (8.14)$$

The conservation of momentum and kinematic relations does not contain any information about the material. Constitutive laws provide this additional information. The most common such law describes a linear elastic material and can be conveniently expressed using a symmetric matrix c_{ij}, called stiffness matrix:

$$\begin{bmatrix} T_{xx} \\ T_{yy} \\ T_{zz} \\ T_{yz} \\ T_{zx} \\ T_{xy} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} & c_{15} & c_{16} \\ & c_{22} & c_{23} & c_{24} & c_{25} & c_{26} \\ & & c_{33} & c_{34} & c_{35} & c_{36} \\ & & & c_{44} & c_{45} & c_{46} \\ & \text{symm.} & & & c_{55} & c_{56} \\ & & & & & c_{66} \end{bmatrix} \cdot \begin{Vmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \\ \varepsilon_{yz} \\ \varepsilon_{zx} \\ \varepsilon_{xy} \end{Vmatrix} \quad (8.15)$$

In the most general case, the matrix c_{ij} has 21 independent elements. When the material has a crystal symmetry, the number of independent constants is reduced. For example, single crystal Si is a common structural material in MEMS with a cubic symmetry. In this case there are only three independent constants:

$$\begin{bmatrix} T_{xx} \\ T_{yy} \\ T_{zz} \\ T_{yz} \\ T_{zx} \\ T_{xy} \end{bmatrix} = \begin{bmatrix} c_{11} & c_{12} & c_{12} & 0 & 0 & 0 \\ & c_{11} & c_{12} & 0 & 0 & 0 \\ & & c_{11} & 0 & 0 & 0 \\ & & & c_{44} & 0 & 0 \\ & \text{symm.} & & & c_{44} & 0 \\ & & & & & c_{44} \end{bmatrix} \cdot \begin{bmatrix} \varepsilon_x \\ \varepsilon_y \\ \varepsilon_z \\ \varepsilon_{yz} \\ \varepsilon_{zx} \\ \varepsilon_{xy} \end{bmatrix} \quad (8.16)$$

If the material is isotropic (amorphous or polycrystalline), the number of independent elastic constants is further reduced to two by the relation $c_{44} = (c_{11} - c_{12})/2$. The elastic constants of several most commonly used materials are listed in Table 8.1 (from [Kittel 1996]).

Additional information on other symmetry classes can be found in [Nye 1960].

TABLE 8.1 Elastic Constants of Several Common Cubic Crystals

Crystal	Stiffness Constants at Room Temperature, 10^{11} N/m^2		
	c_{11}	c_{12}	c_{44}
W	5.233	2.045	1.607
Ta	2.609	1.574	0.818
Cu	1.684	1.214	0.754
Ag	1.249	0.937	0.461
Au	1.923	1.631	0.420
Al	1.608	0.607	0.282
K	0.0370	0.0314	0.0188
Pb	0.495	0.423	0.149
Ni	2.508	1.500	1.235
Pd	2.271	1.761	0.17
Si	1.66	0.639	0.796

Electric Phenomena

In the previous section the laws governing the motion of rigid and deformable bodies were reviewed. The forces entering these equations are often of electromagnetic origin; thus one has to know the distribution of electric and magnetic fields. The electromagnetic field is governed by a set of four coupled equations known as Maxwell's equations. Similarly, to the momentum equations, these can also be postulated in integral form. Here we only give the local form

$$\begin{aligned} \dot{\mathbf{B}} + \nabla \times \mathbf{E} &= \mathbf{0} \\ \nabla \cdot \mathbf{D} &= q^f \\ \nabla \times \mathbf{H} - \dot{\mathbf{D}} &= \mathbf{i} \\ \nabla \cdot \mathbf{B} &= 0 \end{aligned} \tag{8.17}$$

where **E** is the electric field, **D** is the electric displacement, **B** is the magnetic induction, **H** is the magnetic field strength, **i** is the electric current density, and q^f is the free charge volume density. Equations (8.17) require constitutive laws specifying the current density, electric displacement, and magnetic field in terms of electric field and magnetic induction vectors. A linear form of these laws is given by

$$\mathbf{i} = \frac{\mathbf{E}}{\rho_e}, \qquad \mathbf{D} = \varepsilon_0 \mathbf{E} + \mathbf{P}, \qquad \mathbf{B} = \mu_0 \mathbf{H} + \mu_0 \mathbf{M} = \mu_0 \mu_r \mathbf{H} \tag{8.18}$$

where ρ_e is the electrical resistance. The coupling between electrical and mechanical fields can be linear or nonlinear. For example, piezoelectricity is a linear phenomenon describing the generation of electric field as a result of the application of mechanical stress. Electrostriction on the other hand is a second order effect, resulting in the generation of mechanical strain proportional to the square of the electric field. Other effects include piezoresistivity, i.e., a change of the electrical resistance due to mechanical stress. In addition to these material properties, electromechanical coupling can be achieved through direct use of electromagnetic forces (Lorentz force) as is commonly done in conventional electrical machines. Lorentz force per unit volume is given by

$$\mathbf{f}^L = q^f(\mathbf{E} + \mathbf{v} \times \mathbf{B}) \tag{8.19}$$

where q^f is the volume charge density. Equation (8.19) accounts for the forces acting on free charge only. If the fields have strong gradients, the above expression should be modified to include the polarization and magnetization terms [Maugin 1988].

$$\mathbf{f}^{\mathrm{EM}} = q^f \mathbf{E} + \left(\mathbf{i} + \frac{\partial \mathbf{P}}{\partial t}\right) \times \mathbf{B} + \mathbf{P} \cdot \nabla \mathbf{E} + \nabla \mathbf{B} \cdot \mathbf{M} \tag{8.20}$$

Equation (8.19) or (8.20) can be used in the momentum equation (8.10) in place of the body force **f**.

As mentioned earlier, piezoelectricity and piezoresistivity are the other commonly used effects in electromechanical systems. The piezoelectric effect occurs only in materials with certain crystal structure. Common examples include $BaTiO_3$ and lead zirconia titanate (PZT). In the quasi-electrostatic approximation (when the magnetic effects are neglected) there are four variables describing the electromechanical state of the body—electric field **E** and displacement **D**, mechanical stress **T** and strain $\boldsymbol{\varepsilon}$. The constitutive laws of piezoelectricity are given as a set of two matrix equations between the four field variables, relating one mechanical and one electrical variable to the other two in the set

$$\varepsilon_{ij} = s_{ijkl} T_{kl} + d_{ijk} E_k, \qquad D_i = d_{ikl} T_{kl} + \varepsilon_0 \Xi_{ij} E_j \tag{8.21}$$

where s_{ijkl} is the elastic compliance tensor, d_{ijk} is the piezoelastic tensor, Ξ_{ij} is the electric permittivity tensor. If the electric field and the polarization vectors are co-linear, the stress and strain tensors are symmetric, and the number of independent coefficients in s_{ijkl} is reduced from 81 to 21 and for the piezoelastic tensor d_{ijk} from 27 to 18. If further, the piezoelectric is poled in one direction only (for example index 3),

the only nonzero elements are

$$d_{113},\quad d_{223},\quad d_{333},\quad d_{232} = d_{322},\quad d_{131} = d_{313},\quad d_{123} = d_{213}.$$

Numerical values for the coefficients in (8.22) for bulk $BaTiO_3$ crystals can be found in [Zgonik et al. 1994].

8.2 Common Structures in Mechatronic Systems

Microelectromechanical systems (MEMS) traditionally use technology developed for the manufacturing of integrated circuits. As a result, the employed mechanical structures are often planar devices—springs, coils, bridges, or cantilever beams, subjected to in-plane and out-of-plane bending and torsion. Using high aspect ratio reactive ion etching combined with fusion bonding of silicon, it is possible to realize true three-dimensional structures as well. For example Fig. 8.2 shows an SEM micrograph of a complex capacitive force sensor designed to accept glass fibers in an etched v-groove. In this section, we will review the fundamental relationships used in the initial designs of such electromechanical systems.

Beams

Microcantilevers are used in surface micromachined electrostatic switches, as "cantilever tip" for scanning probe microscopy (SPM) and in myriad of sensors, based on vibrating cantilevers. The majority of the surface micromachined beams fall into two cases—cantilever beams and bridges. Figure 8.3 illustrates a two-layer cantilever beam (Fig. 8.3(a)) and a bridge (Fig. 8.3(b)). The elastic force required to produce deflection d at the tip of the cantilever beam, or at the center of the bridge, is given by

$$F^{\text{elast}} = K_{\text{eff}}\, d \tag{8.22}$$

where

$$K_{\text{eff}} = \frac{24(EI)_{\text{eff}}}{(6l_e^3/5) + 6(l-l_e)l_e^2 + 12(l-l_e)^2 l_e + 8(l-l_e)^3} \quad \text{and} \quad K_{\text{eff}} = \frac{360(EI)_{\text{eff}}}{30l^3 - 45ll_e^2 - 5(l_e^4/l) + 3l_e^3} \tag{8.23}$$

are the effective spring constants of the composite beams for cantilever and bridge beams, respectively. The effective stiffness of the beam in both cases can be calculated from

$$(EI)_{\text{eff}} = \frac{E_1 w t_1^3}{12} + \frac{E_2 w t_2^3}{12} + \frac{E_1 E_2 t_1 t_2 w (t_1 + t_2)^2}{4(E_1 t_1 + E_2 t_2)} \tag{8.24}$$

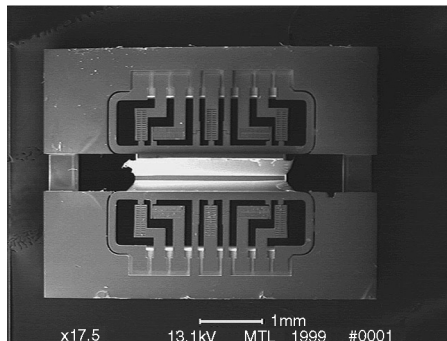



FIGURE 8.2 Capacitive force sensor using 3D micromachining.

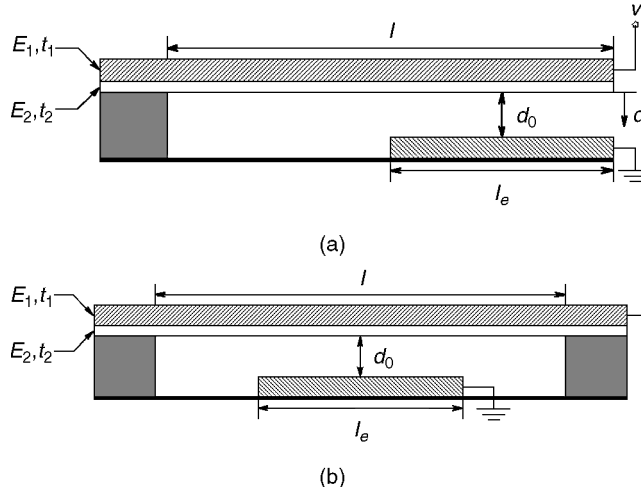



FIGURE 8.3 Surface micromachined beams: (a) Two-layer composite beam with electrostatic actuation; (b) two-layer composite bridge with electrostatic actuation.

where w is the width of the beam, t_1 the thickness of the top beam, t_2 the thickness of insulating layer (silicon oxide, silicon nitride), l the length of the beam, l_e the length of fixed electrode, E_1 the Young's modulus of the top layer, E_2 the Young's modulus of insulating layer.

Torsional Springs

Torsion of beams is used primarily in rotating structures such as micromirrors for optical scanning, or projection displays. The micromirror array developed by Texas Instruments for example uses polycrystalline silicon beams as hinges of the micromirror plate.

The torsion problems can be solved in a closed form for beams with elliptical or triangular cross sections [Mendleson 1968]. In the case of an elliptical cross section, the moment required to produce an angular twist (angle or rotation per unit length of the beam) α [rad/m] is equal to

$$M = \frac{\pi a^3 b^3}{a^2 + b^2} G\alpha \tag{8.25}$$

where G is the elastic shear modulus, and a and b are the lengths of the two semi-axes of the ellipse. The maximum shear stress in this case is

$$\tau^{\max} = \frac{2G\alpha a^2 b}{a^2 + b^2}, \quad a > b \tag{8.26}$$

The torsional stiffness of rectangular cross-section beams can be obtained in terms of infinite power series [Hopkins 1987]. If the cross-section has dimension $a \times b$, $b < a$, the first three term of this series result in an equation similar to (8.25)

$$M = 2KG\alpha, \quad \text{where } K = ab^3\left[\frac{1}{3} - 0.21\frac{b}{a}\left(1 - \frac{b^4}{12a^4}\right)\right]. \tag{8.27}$$

Thin Plates

Pressure sensors are one of the most popular electromechanical transducers. The basic structure used to convert mechanical pressure into electrical signal is a thin plate subjected to a pressure differential. Piezoresistive gauges are used to convert the strain in the membrane into change of resistance, which is

TABLE 8.2 Deflection and Bending Moments of Clamped Plate Under Uniform Load q [Evans 1939]

b/a	$W(x=0, y=0)$	$M_x(x=a/2, y=0)$	$M_y(x=0, y=b/2)$	$M_x(x=0, y=0)$	$M_y(x=0, y=0)$
1	$0.00126qa^4/D$	$-0.0513qa^2$	$-0.0513qa^2$	$0.0231qa^2$	$0.0231qa^2$
1.5	$0.00220qa^4/D$	$-0.0757qa^2$	$-0.0570qa^2$	$0.0368qa^2$	$0.0203qa^2$
2	$0.00254qa^4/D$	$-0.0829qa^2$	$-0.0571qa^2$	$0.0412qa^2$	$0.0158qa^2$
∞	$0.00260qa^4/D$	$-0.0833qa^2$	$-0.0571qa^2$	$0.0417qa^2$	$0.0125qa^2$

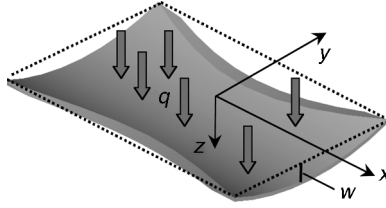



FIGURE 8.4 Thin plate subjected to positive pressure q.

read out using a conventional resistive bridge circuit. The initial pressure sensors were fabricated via anisotropic etching of silicon, which results in a rectangular diaphragm. Figure 8.4 shows a thin-plate, subjected to normal pressure q, resulting in out-of-plane displacement $w(x, y)$. The equilibrium condition for $w(x, y)$ is given by the thin plate theory [Timoshenko 1959]:

$$\frac{\partial^4 w}{\partial x^4} + 2\frac{\partial^4 w}{\partial x^2 \partial y^2} + \frac{\partial^4 w}{\partial y^4} = \frac{q}{D}, \tag{8.28}$$

where $D = Eh^3/12(1-\nu^2)$ is the flexural rigidity, E is the Young's modulus, ν is the Poisson ratio, and h is the thickness of the plate. The edge-moments (moments per unit length of the edge) and the small strains are

$$\begin{aligned} M_x(x, y) &= -D\left(\frac{\partial^2 w}{\partial x^2} - \nu\frac{\partial^2 w}{\partial y^2}\right), & \varepsilon_{xx}(x, y, z) &= -z\frac{\partial^2 w}{\partial x^2} \\ M_y(x, y) &= -D\left(\frac{\partial^2 w}{\partial y^2} - \nu\frac{\partial^2 w}{\partial x^2}\right), & \varepsilon_{yy}(x, y, z) &= -z\frac{\partial^2 w}{\partial y^2} \\ M_{xy}(x, y) &= D(1-\nu)\frac{\partial^2 w}{\partial x \partial y}, & \varepsilon_{xy}(x, y, z) &= -z\frac{\partial^2 w}{\partial x \partial y} \end{aligned} \tag{8.29}$$

Using (8.29), one can calculate the maximum strains occurring at the top and bottom faces of the plate in terms of the edge-moments:

$$\begin{aligned} \varepsilon_{xx}^{\max}(x, y, z) &= \frac{12z}{Eh^3}(M_x - \nu M_y)\big|_{z=h} = \frac{12}{Eh^2}(M_x - \nu M_y) \\ \varepsilon_{yy}^{\max}(x, y, z) &= \frac{12z}{Eh^3}(M_y - \nu M_x)\big|_{z=h} = \frac{12}{Eh^2}(M_y - \nu M_x) \end{aligned} \tag{8.30}$$

In the case of a pressure sensor with a diaphragm subjected to a uniform pressure, the boundary conditions are built-in edges: $w = 0$, $\partial w/\partial x = 0$ at $x = \pm a/2$ and $w = 0$, $\partial w/\partial y = 0$ at $y = \pm b/2$, where the diaphragm has lateral dimensions $a \times b$. The solution of this problem has been obtained by [Evans 1939], showing that the maximum strains are at the center of the edges. The values of the edge-moments and the displacement of the center of plate are listed in Table 8.2.

8.3 Vibration and Modal Analysis

As mentioned earlier, the time response of a continuum structure requires the solution of Eqs. (8.10) with the acceleration terms present. For linear systems this solution can be represented by an infinite superposition of characteristic functions (modes). Associated with each such mode is also a characteristic number (eigenvalue) determining the time response of the mode. The analysis of these modes is called modal analysis and has a central role in the design of resonant cantilever sensors, flapping wings for micro-air-vehicles (MAVs) and micromirrors, used in laser scanners and projection systems. In the case of a cantilever beam, the flexural displacements are described by a fourth-order differential equation

$$\frac{IE}{\rho A}\frac{\partial^4 w(x,t)}{\partial x^4} + \frac{\partial^2 w(x,t)}{\partial t^2} = 0 \qquad (8.31)$$

where I is the moment of inertia, E is the Young's modulus, ρ is the density, and A is the area of the cross section. When the thickness of the cantilever is much smaller than the width, E should be replaced by the reduced Young's modulus $E_1 = E/(1 - \nu^2)$. For a rectangular cross section, (8.31) is reduced to

$$\frac{Eh^2}{12\rho}\frac{\partial^4 w(x,t)}{\partial x^4} + \frac{\partial^2 w(x,t)}{\partial t^2} = 0 \qquad (8.32)$$

where h is the thickness of the beam. The solution of (8.32) can be written in terms of an infinite series of characteristic functions representing the individual vibration modes

$$w = \sum_{i=1}^{\infty} \Phi_i(x)\sin(\omega_i t + \delta_i) \qquad (8.33)$$

where the characteristic functions Φ_i are expressed with the four Rayleigh functions S, T, U, and V:

$$\begin{aligned}
\Phi_i &= a_i S(\lambda_i x) + b_i T(\lambda_i x) + c_i U(\lambda_i x) + d_i V(\lambda_i x) \\
S(x) &= \frac{1}{2}(\cosh x + \cos x), \quad T(x) = \frac{1}{2}(\sinh x + \sin x) \\
U(x) &= \frac{1}{2}(\cosh x - \cos x), \quad V(x) = \frac{1}{2}(\sinh x - \sin x), \quad \lambda_i^4 = \omega_i^2 \frac{\rho A}{IE}
\end{aligned} \qquad (8.34)$$

The coefficients a_i, b_i, c_i, d_i, ω_i, and δ_i are determined from the boundary and initial conditions of (8.34). For a cantilever beam with a fixed end at $x = 0$ and a free end at $x = L$, the boundary conditions are

$$\begin{aligned}
w(0,t) &= 0, \quad \frac{\partial^2 w(L,t)}{\partial x^2} = 0 \\
\frac{\partial w(0,t)}{\partial x} &= 0, \quad \frac{\partial^3 w(L,t)}{\partial x^3} = 0
\end{aligned} \qquad (8.35)$$

Since (8.35) are to be satisfied by each of the functions Φ_i, it follows that $a_i = 0$, $b_i = 0$ and

$$\cosh(\lambda_i L)\cos(\lambda_i L) = -1 \qquad (8.36)$$

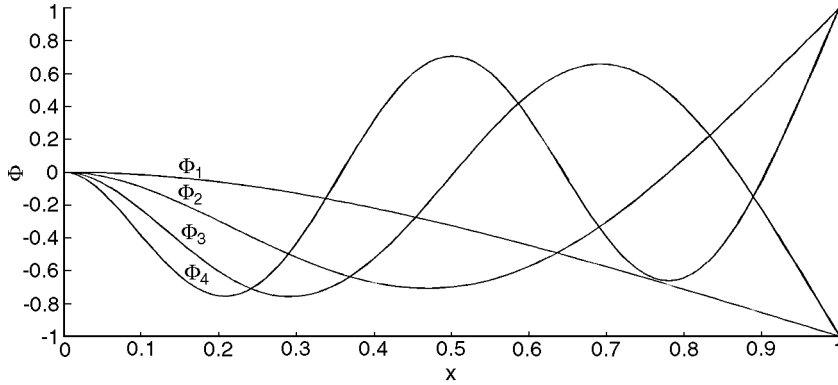



FIGURE 8.5 First four vibration modes of a cantilever beam.

From this transcendental equation the λ_i's and the circular frequencies ω_i are determined [Butt et al. 1995].

$$\lambda_i L \cong \frac{(2i-1)\pi}{2}, \qquad \omega_i = \frac{(2i-1)^2\pi^2}{4L^2}\sqrt{\frac{IE}{\rho A}} = \frac{(2i-1)^2\pi^2}{4L^2}\sqrt{\frac{Eh^2}{12\rho}} \tag{8.37}$$

Figure 8.5 shows the first four vibrational modes of the cantilever. An important result of the modal analysis is the calculation of the amplitude of thermal vibrations of cantilevers. As the size of the cantilevers is reduced to nanometer scale, the energy of random thermal excitations becomes comparable with the energy of the individual vibration modes. This effect leads to a thermal noise in nanocantilevers. Using the equipartition theorem [Butt et al. 1995] showed that the root mean square of the amplitude of the tip of such cantilever is

$$\sqrt{\hat{z}^2} = \sqrt{\frac{kT}{K}} = \frac{0.64\ \text{Å}}{\sqrt{K}}, \quad K = \frac{Ewh^3}{4L^2} \tag{8.38}$$

Similar analysis can be performed on vibrations of thin plates such as micromirrors. The free lateral vibrations of such a plate are described by

$$\frac{\partial^4 w(x,y,t)}{\partial x^4} + 2\frac{\partial^4 w(x,y,t)}{\partial x^2 \partial y^2} + \frac{\partial^4 w(x,y,t)}{\partial y^4} = -\frac{\rho h}{D}\frac{\partial^2 w(x,y,t)}{\partial t^2} \tag{8.39}$$

The interested reader is referred to [Timoshenko 1959] for further details on vibrations of plates.

8.4 Buckling Analysis

Structural instability can occur due to material failure, e.g., plastic flow or fracture, or it can also occur due to large changes in the geometry of the structure (e.g., buckling, wrinkling, or collapse). The latter is the scope of this section. When short columns are subjected to a compressive load, the stress in the cross section is considered uniform. Thus for short columns, failure will occur when the plastic yield stress of the material is reached. In the case of long and slender beams under compression, due to manufacturing imperfections, the applied load or the column will have some eccentricity. As a result the force will develop a bending moment proportional to the eccentricity, resulting in additional lateral deflection. While for small loads the lateral displacement will reach equilibrium, above certain critical

TABLE 8.3 Critical Load Coefficients

	End Conditions		
	one end built-in, other free	both ends built-in	pin-joints at both ends
K coefficient	1/4	4	1

load, the beam will be unable to withstand the bending moment and will collapse. Consider the beam in Fig. 8.5, subjected to load F with eccentricity e, resulting in lateral displacement of the tip δ. According to the beam bending equation

$$EI\frac{\partial^2 w}{\partial x^2} = M = F(\delta + e + w) \tag{8.40}$$

where the boundary conditions are $w(0) = 0$, $\partial w/\partial x\,|_{x=0} = 0$. The corresponding solution is

$$w = (e + \delta)[1 - \cos(\sqrt{IE/F}x)] \tag{8.41}$$

From $w(L) = \delta$ one has $\delta = e(1/\cos kL - 1)$, where $k = \sqrt{\mathrm{IE}/F}$. This solution looses stability when δ grows out of bound, i.e., when $\cos kL = 0$, or $kL = (2n + 1)\pi/2$. From this condition the smallest critical load is

$$F^{cr} = \pi^2 IE/4L^2 \tag{8.42}$$

The above analysis and Eq. (8.42) were developed by Euler. Similar conditions can be derived for other types of beam supports. A general formula for the critical load can be written as

$$F^{cr} = K\pi^2 IE/L^2 \tag{8.43}$$

where several values of the coefficient K are given in Table 8.3.

8.5 Transducers

Transducers are devices capable of converting one type of energy into another. If the output energy is mechanical work the transducer is called an actuator. The rest of the transducers are called sensors, although in most cases, a mechanical transducer can also be a sensor and vice versa. For example the capacitive transducer can be used as an actuator or position sensor. In this section the most common actuators used in micromechatronics are reviewed.

Electrostatic Transducers

The electrostatic transducers fall into two main categories—parallel plate electrodes and interdigitated comb electrodes. In applications where relatively large capacitance change or force is required, the parallel plate configuration is preferred. Conversely, larger displacements with linear force/displacement characteristics can be achieved with comb drives at the expense of reduced force. Parallel plate actuators are used in electrostatic micro-switches as illustrated in Fig. 8.1. In this case the electrodes form a parallel plate capacitor and the force is described by

$$F_{\text{elec}} = \frac{A\varepsilon_0\varepsilon_r^2 V^2}{2[t_2 + \varepsilon_r(d_0 - d)]^2} \tag{8.44}$$

where A is the area of overlap between the two electrodes; t_2 is the thickness of insulating layer (silicon dioxide, silicon nitride); l_e is the length of fixed electrode; ε_r is the relative permittivity of insulating layer; V is the applied voltage; d_0 is the initial separation between the capacitor plates; and d is downward

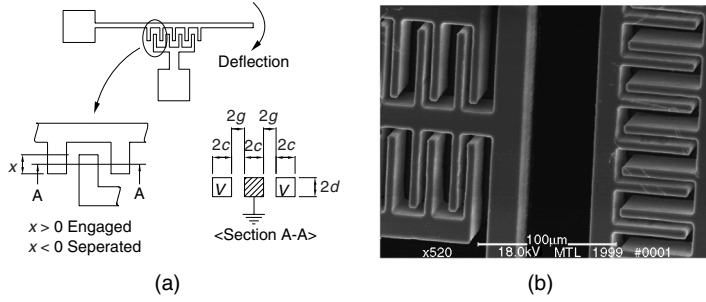



FIGURE 8.6 Lateral comb transducers: (a) Dimensions; (b) two orthogonal Si combs.

deflection of the beam. The minimum voltage required to close the gap of a cantilever actuator is known as the threshold voltage [Petersen 1978], and can be approximated as

$$V^{th} \approx \sqrt{\frac{18(IE)_{eff}\, d_0^3}{5\varepsilon_0 L^4 w}} \tag{8.45}$$

where $(IE)_{eff}$ is given by (8.24).

Comb drives also fall in two categories: symmetric and asymmetric. Symmetric comb drive is shown in Fig. 8.6(a). In this configuration the gaps between the individual fingers are equal. Figure 8.6(b) shows a pair of asymmetric comb capacitors, used in the force sensor shown in Fig. 8.2 [Enikov 2000a]. In any case, the force generated between the fingers is equal to the derivative of the total electrostatic energy with respect to the displacement

$$F^{el} = \frac{n}{2}\frac{\partial C}{\partial x}V^2 \tag{8.46}$$

where n is the number of fingers. Several authors have given approximate expressions for (8.46). One of the most accurate calculations of the force between the pair of fingers shown in Fig. 8.6(a) is given by [Johnson et al. 1995] using Schwartz transforms

$$F^{el} = \begin{cases} \dfrac{\varepsilon_0 V^2}{\pi}\left\{\ln\left[\left(\left(\dfrac{c}{g}+1\right)^2-1\right)\left(1+\dfrac{2g}{c}\right)^{1+c/g}\right]+\dfrac{\pi d}{g}-\dfrac{c+g}{x}\right\}, & x > \Delta_+ \text{ (engaged)} \\ -\dfrac{\varepsilon_0 V^2}{\pi}\left\{\dfrac{2(c+g)}{x}\right\}, & x < -\Delta_- \text{ (separated)} \end{cases} \tag{8.47}$$

In the transition region $x \in [-\Delta_-; \Delta_+]$, $\Delta_{+,-} \approx 2g$, the force can be approximated with a tangential line between the two branches described by (8.47).

Electromagnetic Transducers

Electromagnetic force has also been used extensively. It can be generated via planar coil as illustrated in Fig. 8.7. The cantilever and often the coils are made of soft ferromagnetic material. Using an equivalent magnetic circuit model, the magnetic force acting on the top cantilever can be estimated as

$$F_{mag} = \frac{2n^2 I^2 (2A_2 + A_1)}{\mu_0 A_1 A_2 (2R_1 + R_2)^2} \tag{8.48}$$

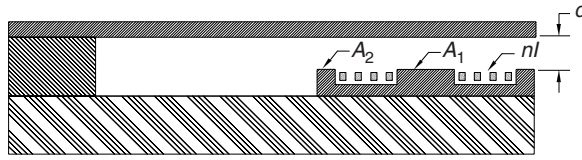



FIGURE 8.7 Electromagnetic actuation.

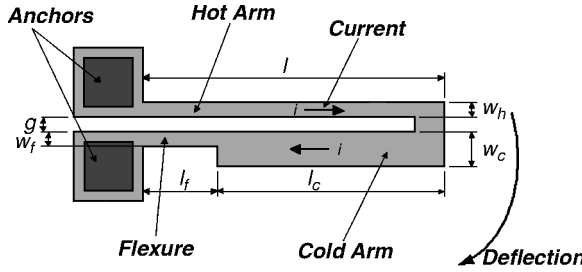



FIGURE 8.8 Lateral thermal actuator.

where

$$R_1 = \frac{d}{\mu_0 A_1} + \frac{h_1}{\mu_0 \mu_r A_1}, \qquad R_2 = \frac{d}{\mu_0 A_2} + \frac{h_1}{\mu_0 \mu_r A_2} + \frac{h_2}{\mu_0 \mu_r A_b} \tag{8.49}$$

are the reluctances; h_1 and h_2 are the flux-path lengths inside the top and bottom permalloy layers.

Thermal Actuators

Thermal actuators have been investigated for positioning of micromirrors [Liew et al. 2000], and micro-switch actuation [Wood et al. 1998]. This actuator consists of two arms with different cross sections (see Fig. 8.8). When current is passed through the two arms, the higher current density occurs in the smaller cross-section beam and thus generates more heat per unit volume. The displacement is a result of the temperature differential induced in the two arms. For the actuator shown in Fig. 8.8, an approximate model for the deflection of the tip δ can be developed using the theory of thermal bimorphs [Faupel 1981]

$$\delta \approx \frac{3l^2(T^{\text{hot}}\alpha(T^{\text{hot}}) - T^{\text{cold}}\alpha(T^{\text{cold}}))}{4(w_h + w_f)} \tag{8.50}$$

where T^{hot} and T^{cold} are the average temperatures of the hot and cold arms and $\alpha(T)$ is the temperature dependent thermal expansion coefficient. A more detailed analysis including the temperature distribution in the arms can be found in [Huang et al. 1999].

Electroactive Polymer Actuators

Electroactive polymer-metal composites (EAPs) are promising multi-functional materials with extremely reach physics. Recent interest towards these materials is driven by their unique ability to undergo large deformations under very low driving voltages as well as their low mass and high fracture toughness. For comparison, Table 8.4 lists several characteristic properties of EAPs and other piezoelectric ceramics.

EAPs are being tested for use in flapping-wing micro-air-vehicles (MAVs) [Rohani 1999], underwater swimming robots [Laurent 2001], and biomedical applications [Oguro 2000]. An EAP actuator consists

TABLE 8.4 Comparative Properties of EAPs, Shape Memory Alloy, and Piezoceramic Actuators

Characteristic Property	EAP	Shape Memory Alloy	Piezoelectric Ceramics
Achievable strain	more than 10%	up to 8%	up to 0.3%
Young's modulus (GPa)	0.114 (wet)	75	89
Tensile strength (MPa)	34 (wet)	850	76
Response time	msec–min	sec–min	μsec–sec
Mass density (g/cm^3)	2.0	6.5	7.5
Actuation voltage	1–10 V	N/A	50–1000 V

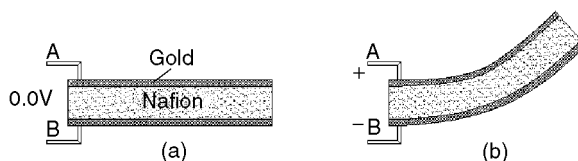



FIGURE 8.9 Polymer metal composite actuator.

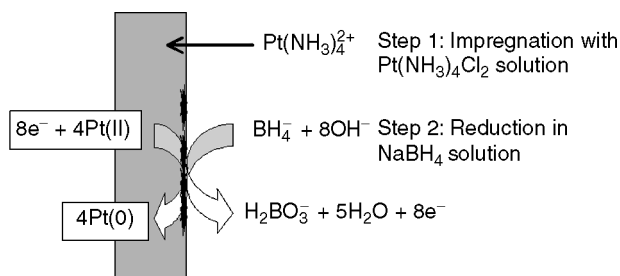
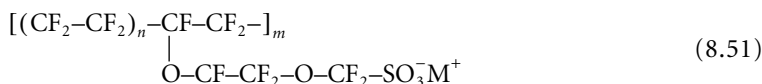



FIGURE 8.10 Two-step Pt plating process.

of an ion-exchange membrane covered with a conductive layer as illustrated in Fig. 8.9(a). Upon application of a potential difference at points A and B the composite bends towards the anodic side as shown in Fig. 8.9(b). Among the numerous ion-exchange polymers, perfluorinated sulfonic acid (Nafion Du Pont, USA) and perfluorinated carboxylic acid (Flemion, Asahi, Japan) are the most commonly used in actuator applications. The chemical formula of a unit chain of Nafion is

$$\begin{array}{l}[(CF_2\text{–}CF_2)_n\text{–}CF\text{–}CF_2\text{–}]_m \\ \qquad\qquad\quad | \\ \qquad\qquad\quad O\text{–}CF\text{–}CF_2\text{–}O\text{–}CF_2\text{–}SO_3^-M^+\end{array} \tag{8.51}$$

where M^+ is the counterion (H^+, Na^+, Li^+, …). The ionic clusters are attached to side chains, which according to transmission electron microscopy (TEM) studies, segregate in hydrophilic nano-clusters with diameters ranging from 10 to 50 Å [Xue 1989]. In 1982, Gierke proposed a structural model [Gireke 1982] according to which, the clusters are interconnected via narrow channels. The size and distribution of these channels determine the transport properties of the membrane and thus the mechanical response.

Metal-polymer composites can be produced by vapor or electrochemical deposition of metal over the surface of the membrane. The electrochemical platinization method [Fedkiw 1992], used by the author, is based on the ion-exchange properties of the Nafion. The method consists of two steps: step one—ion exchange of the protons H^+ with metal cations (e.g., Pt^{2+}); step two—chemical reduction of the Pt^{2+} ions in the membrane to metallic Pt using $NaBH_4$ solution. These steps are outlined in Fig. 8.10 and an SEM microphotograph of the resulting composite is shown in Fig. 8.11. The electrode surfaces are approximately 0.8 μm thick Pt deposits. Repeating the above steps several times results in dendritic growth of the electrodes into the polymer matrix [Oguro 1999] and has been shown to improve the actuation efficiency.

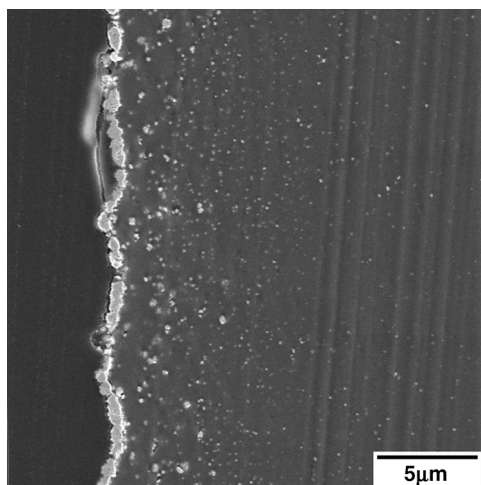



FIGURE 8.11 Nafion membrane with Pt electrode.

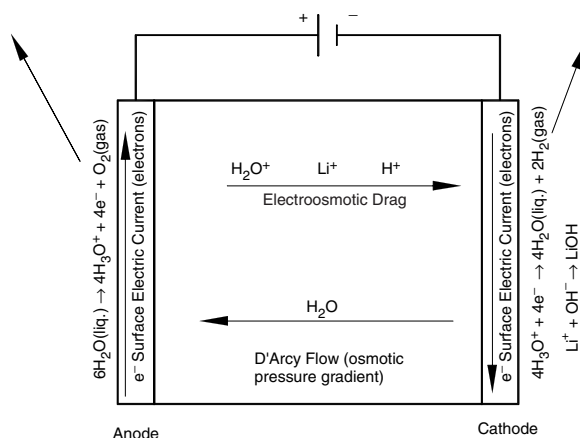



FIGURE 8.12 Ion transort in nafion.

The deformation of the polymer-metal composite can be attributed to several phenomena, the dominant one being differential swelling of the membrane due to internal osmotic pressure gradients [Eikerling 1998]. A schematic representation of the ionic processes taking place inside the polymer is shown in Fig. 8.12. Under the application of external electric field a flux of cations and hydroxonium ions is generated towards the cathode. At the cathode the ions pick up an electron and produce hydrogen and free water molecules. On the anodic side, the water molecules dissociate producing oxygen and hydroxonium ions. This redistribution of water within the membrane creates local expansion/contraction of the polymer matrix. Mathematically, the deformation can be described by introducing an additional strain (eigen strain) term in the expression of the total strain. Thus the total strain has two additive parts: elastic deformation of the polymer network due to external forces (mechanical, electrical) and chemical strain proportional to the compositional variables

$$\varepsilon_{ij} = \varepsilon_{ij}^{\text{elast}} + \rho_0 \sum_s \frac{\bar{V}^s}{3M^s}(c^s - c_0^s)\delta_{ij} \tag{8.52}$$

where c^s are the mass fractions, $\bar{V}^s$ are the partial molar volumes, M^s are the molar masses, and the index 0 refers to the initial value of a variable. Complete mathematical description of the polymer actuator requires the solution of mass transport (diffusion) equation, momentum balance, and Poisson equation for potential distribution, the discussion of which is beyond the scope of this book. An interesting consequence of the addition of the chemical strain in (8.46) is the explicit appearance of the pressure term in the electrochemical potential driving the diffusion. The total mass diffusion flux will have a component proportional to the negative gradient of the pressure, which for the case of water, will result in a relaxation phenomena observed experimentally. The total flux of component s is then given by

$$J^s = -\frac{\rho c^s W^s}{M^s}\nabla(\mu^{os}(T) + p\bar{V}^s + RT\ln(fc^s) + z^s\Phi) \tag{8.53}$$

where W^s is the mobility of component *s*, z^s is the valence of component *s*, *p* is the pressure, f^s is the activity coefficient, and Φ is the electric potential. We have omitted the cross-coupling terms that would appear in a fully coupled Onsager-type formulation. Interested readers are referred to [Enikov 2000b] and the references therein for further details.

8.6 Future Trends

The future MEMS are likely to be more heterogeneous in terms of materials and structures. Bio-MEMS for example, require use of nontoxic, noncorrosive materials, which is not a severe concern in standard IC components. Already departure from the traditional Si-based MEMS can be seen in the areas of optical MEMS using wide band-gap materials, nonlinear electro-optical polymers, and ceramics. As pointed earlier, the submicron size of the cantilever-based sensors brings the thermal noise issues in mechanical structures. Further reduction in size will require molecular statistic description of the interaction forces. For example, carbon nanotubes placed on highly oriented pyrolytic graphite (HOPG) experience increased adhesion force when aligned with the underlying graphite lattice [Falvo et al. 2000]. The future mechatronic systems are likely to become an interface between the macro and nano domains.

References

Butt, H., Jaschke, M., "Calculation of thermal noise in atomic force microscopy," *Nanotechnology,* **6**, pp. 1–7, 1995.

Eikerling, M., Kharkats, Y.I., Kornyshev, A.A., Volfkovich, Y.M., "Phenomenological theory of electro-osmotic effect and water management in polymer proton-conducting membranes," *Journal of the Electrochemical Society,* **145**(8), pp. 2684–2698, 1998.

Evans, T.H., *Journal of Applied Mechanics,* **6**, p. A-7, 1939.

Enikov, E.T., Nelson, B., "Three dimensional microfabrication for multi-degree of freedom capacitive force sensor using fiber chip coupling," *J. Micromech. Microeng.,* **10**, pp. 492–497, 2000.

Enikov, E.T., Nelson, B.J., "Electrotransport and deformation model of ion exhcange membrane based actuators," in *Smart Structures and Materials 2000,* Newport Beach, CA, SPIE vol. 3987, March, 2000.

Falvo, M.R., Steele, J., Taylor, R.M., Superfine, R., "Gearlike rolling motion mediated by commensurate contact: carbon nanotubes on HOPG," *Physical Review B,* **62**(6), pp. 665–667, 2000.

Faupel, J.H., Fisher, F.E., *Engineering Design: A Synthesis of Stress Analysis and Materials Engineering,* 2nd Ed., Wiley & Sons, New York, 1981.

Liu, R., Her, W.H., Fedkiw, P.S., "*In situ* electrode formation on a nafion membrane by chemical platinization," *Journal of the Electrochemical Society,* **139**(1), pp. 15–23, 1990.

Gierke, T.D., Hsu, W.S., "The cluster-network model of ion clusturing in perfluorosulfonated membranes," in *Perfluorinated Ionomer Membranes,* A. Eisenberg and H.L. Yeager, Eds., vol. 180, American Chemical Society, 1982.

Johnson et al., "Electrophysics of micromechanical comb actuators," *Journal of Microelectromechanical Systems,* **4**(1), pp. 49–59, 1995.

Hopkins, *Design Analysis of Shafts and Beams,* 2nd Ed., Malabar, FL: RE Kreiger, 1987.

Huang, Q.A., Lee, N.K.S., "Analysis and design of polysilcon thermal flexture actuator," *Journal of Micromechanics and Microengineering,* **9**, pp. 64–70, 1999.

Kittel, Ch., *Introduction to Solid State Physics,* John Wiley & Sons, Inc., New York, 1996.

Laurent, G., Piat, E., "High efficiency swimming microrobot using ionic polymer metal composite actuators," to appear in 2001.

Liew, L. et al., "Modeling of thermal actuator in a bulk micromachined CMOS micromirror," *Microelectronics Journal,* **31**(9–10), pp. 791–790, 2000.

Maugin, G., *Continuum Mechanics of Electromagnetic Solids,* Elsevier, Amsterdam, The Netherlands, 1988.

Mendelson, *Plasticity: Theory and Application,* Macmillan, New York, 1968.

Nye, J.F., *Physical Properties of Crystals,* Oxford University Press, London, 1960.

Onishi, K., Sewa, Sh., Asaka, K., Fujiwara, N., Oguro, K., "Bending response of polymer electrolyte actuator," in *Smart Structures and Materials 2000,* Newport Beach, CA, SPIE vol. 3987, March, 2000.

Peterson, "Dynamic micromechanics on silicon: techniques and devices," *IEEE,* 1978.

Rohani, M.R., Hicks, G.R., "Multidisciplinary design and prototype of a micro air vehicle," *Journal of Aircraft,* **36**(1), p. 237, 1999.

Timoshenko, S., Woinowsky-Krieger, S., *Theory of Plates and Shells*, McGraw-Hill, New York, 1959.

Wood, R. et al., "MEMS microrelays," *Mechatronics,* **8**, pp. 535–547, 1998.

Xue, T., Trent, Y.S., Osseo-Asare, K., "Characterization of nafion membranes by transmision electron microscopy," *Journal of Membrane Science,* **45**, p. 261, 1989.

Zgonik et al., "Dielectric, elastic, piezoelectric, electro-optic and elasto-optic tensors of $BaTiO_3$ crystals," *Physical Review B,* **50**(9), p. 5841, 1994.

9

Modeling of Mechanical Systems for Mechatronics Applications

Raul G. Longoria
The University of Texas at Austin

9.1 Introduction

Mechatronics applications are distinguished by controlled motion of mechanical systems coupled to actuators and sensors. Modeling plays a role in understanding how the properties and performance of mechanical components and systems affect the overall mechatronic system design. This chapter reviews methods for modeling systems of interconnected mechanical components, initially restricting the

application to basic translational and rotational elements, which characterize a wide class of mechatronic applications. The underlying basis of mechanical motion (kinematics) is presumed known and not reviewed here, with more discussion and emphasis placed on a system dynamics perspective. More advanced applications requiring two- or three-dimensional motion is presented in section 9.6.

Mechanical systems can be conceptualized as rigid and/or elastic bodies that may move relative to one another, depending on how they are interconnected by components such as joints, dampers, and other passive devices. This chapter focuses on those systems that can be represented using lumped-parameter descriptions, wherein bodies are treated as rigid and no dependence on spatial extent need be considered in the elastic effects. The modeling of mechanical systems in general has reached a fairly high level of maturity, being based on classical methods rooted in the Newtonian laws of motion. One benefits from the extensive and overwhelming knowledge base developed to deal with problems ranging from basic mass-spring systems to complex multibody systems. While the underlying physics are well understood, there exist many different means and ways to arrive at an end result. This can be especially true when the need arises to model a multibody system, which requires a considerable investment in methods for formulating and solving equations of motion. Those applications are not within the scope of this chapter, and the immediate focus is on modeling basic and moderately complex systems that may be of primary interest to a mechatronic system designer/analyst.

9.2 Mechanical System Modeling in Mechatronic Systems

Initial steps in modeling any physical system include defining a system boundary, and identifying how basic components can be partitioned and then put back together. In mechanical systems, these analyses can often be facilitated by identifying points in a system that have a distinct velocity. For purposes of analysis, active forces and moments are "applied" at these points, which could represent energetic interactions at a system boundary. These forces and moments are typically applied by actuators but might represent other loads applied by the environment.

A mechanical component modeled as a point mass or rigid body is readily identified by its velocity, and depending on the number of bodies and complexity of motion there is a need to introduce a coordinate system to formally describe the kinematics (e.g., see [12] or [15]). Through a kinematic analysis, additional (relative) velocities can be identified that indicate the connection with and motion of additional mechanical components such as springs, dampers, and/or actuators. The interconnection of mechanical components can generally have a dependence on geometry. Indeed, it is dependence of mechanical systems on geometry that complicates analysis in many cases and requires special consideration, especially when handling complex systems.

A preliminary description of a mechanical system should also account for any constraints on the motional states, which may be functions of time or of the states themselves. The dynamics of mechanical systems depends, in many practical cases, on the effect of constraints. Quantifying and accounting for constraints is of paramount importance, especially in multibody dynamics, and there are different schools of thought on how to develop models. Ultimately, the decision on a particular approach depends on the application needs as well as on personal preference.

It turns out that a fairly large class of systems can be understood and modeled by first understanding basic one-dimensional translation and fixed-axis rotation. These systems can be modeled using methods consistent with those used to study other systems, such as those of an electric or hydraulic type. Furthermore, building interconnected mechatronic system models is facilitated, and it is usually easier for a system analyst to conceptualize and analyze these models.

In summary, once an understanding of (a) the system components and their interconnections (including dependence on geometry), (b) applied forces/torques, and (c) the role of constraints, is developed, dynamic equations fundamentally due to Newton can be formulated. The rest of this section introduces the selection of physical variables consistent with a power flow and energy-based approach to modeling basic mechanical translational and rotational systems. In doing so, a bond graph approach [28,3,17] is introduced for developing models of mechanical systems. This provides a basis for introducing the

concept of causality, which captures the input–output relationship between power-conveying variables in a system. The bond graph approach provides a way to understand and mathematically model basic as well as complex mechanical systems that is consistent with other energetic domains (electric, electromechanical, thermal, fluid, chemical, etc.).

Physical Variables and Power Bonds

Power and Energy Basis

One way to consistently partition and connect subsystem models is by using power and energy variables to quantify the system interaction, as illustrated for a mechanical system in Fig. 9.1(a). In this figure, one **port** is shown at which power flow is given by the product of force and velocity, $F \cdot V$, and another for which power is the product of torque and angular velocity, $T \cdot \omega$. These power-conjugate variables (i.e., those whose product yields power) along with those that would be used for electrical and hydraulic energy domains are summarized in Table 9.1. Similar effort (e) and flow (f) variables can be identified for other energy domains of interest (e.g., thermal, magnetic, chemical). This basis assures energetically correct models, and provides a consistent way to connect system elements together.

In modeling energetic systems, energy continuity serves as a basis to classify and to quantify systems. Paynter [28] shows how the energy continuity equation, together with a carefully defined port concept, provides a basis for a generalized modeling framework that eventually leads to a bond graph approach. Paynter's reticulated equation of energy continuity,

$$-\sum_{i=1}^{l} P_i = \sum_{j=1}^{m} \frac{dE_j}{dt} + \sum_{k=1}^{n} (P_d)_k \tag{9.1}$$

concisely identifies the l distinct flows of power, P_i, m distinct stores of energy, E_j, and the n distinct dissipators of energy, P_d. Modeling seeks to refine the descriptions from this point. For example, in a simple mass–spring–damper system, the mass and spring store energy, a damper dissipates energy, and

TABLE 9.1 Power and Energy Variables for Mechanical Systems

Energy Domain	Effort, e	Flow, f	Power, P
General	e	f	$e \cdot f$ [W]
Translational	Force, F [N]	Velocity, V [m/sec]	$F \cdot V$ [N m/sec, W]
Rotational	Torque, T or τ [N m]	Angular velocity, ω [rad/sec]	$T \cdot \omega$ [N m/sec, W]
Electrical	Voltage, v [V]	Current, i [A]	$v \cdot i$ [W]
Hydraulic	Pressure, P [Pa]	Volumetric flowrate, Q [m^3/sec]	$P \cdot Q$ [W]

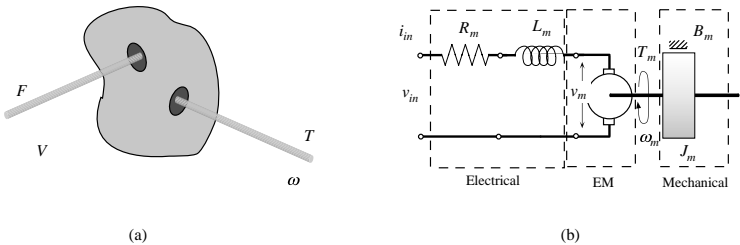



FIGURE 9.1 Basic interconnection of systems using power variables.

the interconnection of these elements would describe how power flows between them. Some of the details for accomplishing these modeling steps are presented in later sections.

One way to proceed is to define and categorize types of system elements based on the reticulated energy continuity Eq. (9.1). For example, consider a system made up only of rigid bodies as energy stores (in particular of kinetic energy) for which $P_d = 0$ (we can add these later), and in general there can be l ports that could bring energy into this purely (kinetic)energy-storing system which has m distinct ways to put energy into the rigid bodies. This is a very general concept, consistent with many other ways to model physical systems. Howevever, it is this foundation that provides for a generalized way to model and integrate different types of energetic systems.

The schematic of a permanent-magnet dc (PMDC) motor shown in Fig. 9.1(b) illustrates how power variables would be used to identify inteconnection points. This example also serves to identify the need for modeling mechanisms, such as the electromechanical (EM) interaction, that can represent the exchange of energy between two parts of a system. This model represents a simplified relationship between electrical power flow, $v \cdot i$, and mechanical power flow, $T \cdot \omega$, which forms the basis for a motor model. Further, this is an ideal power-conserving relationship that would only contain the power flows in the energy continuity equation; there are no stores or dissipators. Additional physical effects would be included later.

Power and Signal Flow

In a bond graph formulation of the PMDC motor, a **power bond** is used to identify flow of power. Power bonds quantify power flow via an effort-flow pair, which can label the bonds as shown in Fig. 9.2(a) (convention calls for the effort to take the position above for any orientation of bond). This is a **word bond graph** model, a form used to identify the essential components in a complex system model. At this stage in a model, only the interactions of multiport systems are captured in a general fashion. Adding half-arrows on power bonds defines a power flow direction between two systems (positive in the direction of the arrow). **Signal bonds**, used in control system diagrams, have full-arrows and can be used in bond graph models to indicate interactions that convey only information (or negligible power) between multiports. For example, the word bond graph in Fig. 9.2(b) shows a signal from the mechanical block to indicate an ideal measurement transferred to a controller as a pure signal. The controller has both signal and power flow signals, closing the loop with the electrical side of the model. These conceptual diagrams are useful for understanding and communicating the system interconnections but are not complete or adequate for quantifying system performance.

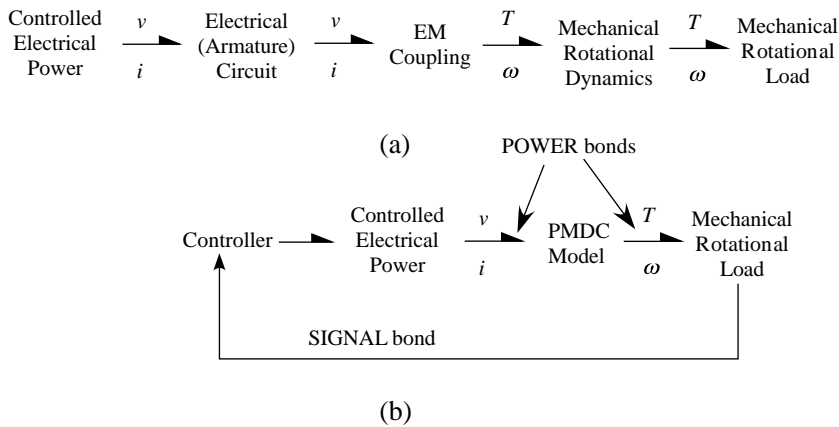



FIGURE 9.2 Power-based bond graph models: (a) PMDC motor word bond graph, (b) PMDC motor word bond graph with controller.

While it is convenient to use power and energy in formulating system models for mechanical systems, a **motional** basis is critical for identifying interconnections and when formulating quantifiable mathematical models. For many mechanical, translational, and rotational systems, it is sufficient to rely on basic one-dimensional motion and relative motion concepts to identify the interrelation between many types of practical components. Identifying network-like structure in these systems has been the basis for building electrical analogies for some time. These methods, as well as signal-flow analysis techniques, are not presented here but are the method of choice in some approaches to system dynamics [33]. Bond graph models are presented, and it will be shown in later sections how these are consistent even with more complex mechanical system formulations of three-dimensional dynamics as well as with the use of Lagrangian models.

Need for Motional Basis

In modeling mechanical translational or rotational systems, it is important to identify how the configuration changes, and a coordinate system should be defined and the effect of geometric changes identified. It is assumed that the reader is familiar with these basic concepts [12]. Usually a reference configuration is defined from which coordinates can be based. This is essential even for simple one-dimensional translation or fixed-axis rotation. The minumum number of geometrically independent coordinates required to describe the configuration of a system is traditionally defined as the **degrees of freedom**. Constraints should be identified and can be used to choose the most convenient set of coordinates for description of the system. We distinguish between degrees of freedom and the minimum number of **dynamic state variables** that might be required to describe a system. These may be related, but they are not necessarily the same variables or the same in number (e.g., a second-order system has two states but is also referred to as a single degree of freedom system).

An excellent illustration of the relevance of degrees of freedom, constraints, and the role these concepts play in modeling and realizing a practical system is shown in Fig. 9.3. This illustration (adapted from Matschinsky [22]) shows four different ways to configure a wheel suspension. Case (a), which also forms the basis for a 1/4-car model clearly has only one degree of freedom. The same is true for cases (b) and (c), although there are constraints that reduce the number of coordinates to just one in each of these designs. Finally, the rigid beam axle shows how this must have two degrees of freedom in vertical and rotational motion of the beam to achieve at least one degree of freedom at each wheel.

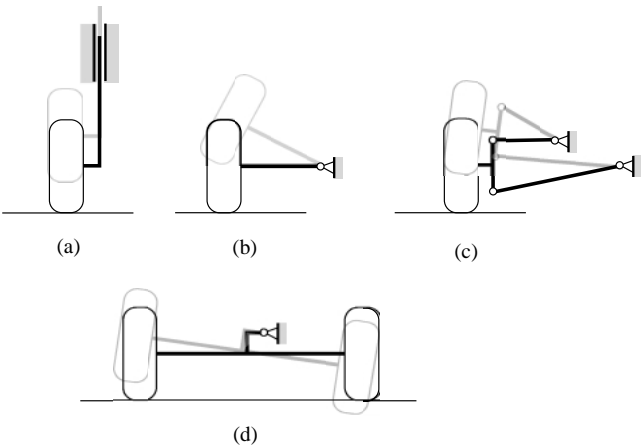



FIGURE 9.3 Wheel suspensions: (a) vertical travel only, 1 DOF; (b) swing-axle with vertical and lateral travel, 1 DOF; (c) four-bar linkage design, constrained motion, 1 DOF; (d) rigid beam axle, two wheels, vertical, and rotation travel, 2 DOF.

Interconnection of Components

In this chapter, we will use bond graphs to model mechanical systems. Like other graph representations used in system dynamics [33] and multibody system analysis [30,39], bond graphs require an understanding of basic model elements used to represent a system. However, once understood, graph methods provide a systematic method for representing the interconnection of multi-energetic system elements. In addition, bond graphs are unique in that they are not linear graph formulations: power bonds replace branches, multiports replace nodes [28]. In addition, they include a systematic approach for computational causality.

Recall that a single line represents power flow, and a half-arrow is used to designate positive power flow direction. Nodes in a linear graph represent across variables (e.g., velocity, voltage, flowrate); however, the multiport in a bond graph represents a system element that has a physical function defined by an energetic basis. System model elements that represent masses, springs, and other components are discussed in the next section. Two model elements that play a crucial role in describing how model elements are interconnected are the 1-junction and 0-junction. These are ideal (power-conserving) multiport elements that can represent specific physical relations in a system that are useful in interconnecting other model elements.

A point in a mechanical system that has a distinct velocity is represented by a 1-junction. When one or more model elements (e.g., a mass) have the same velocity as a given 1-junction, this is indicated by connecting them to the 1-junction with a power bond. Because the 1-junction is constrained to conserve power, it can be shown that efforts (forces, torques) on all the connected bonds must sum to zero; i.e., Â$e_i = 0$. This is illustrated in Fig. 9.4(a). The 1-junction enforces kinematic compatibility and introduces a way to graphically express force summation! The example in Fig. 9.4(b) shows three systems (the blocks labeled 1, 2, and 3) connected to a point of common velocity. In the bond graph, the three systems would be connected by a 1-junction. Note that sign convention is incorporated into the sense of the power arrow.

For the purpose of analogy with electrical systems, the 1-junction can be thought of as a series electrical connection. In this way, elements connected to the 1-junction all have the same current (a flow variable) and the effort summation implied in the 1-junction conveys the Kirchhoff voltage law. In mechanical systems, 1-junctions may represent points in a system that represent the velocity of a mass, and the effort summation is a statement of Newton's law (in D'Alembert form), Â$F - \dot{p} = 0$.

Figure 9.4 illustrates how components with common velocity are interconnected. Many physical components may be interconnected by virtue of a common effort (i.e., force or torque) or 0-junction. For example, two springs connected serially deflect and their ends have distinct rates of compression/extension; however, they have the same force across their ends (ideal, massless springs). System components that have this type of relationship are graphically represented using a 0-junction. The basic 0-junction definition is shown in Fig. 9.5(a). Zero junctions are especially helpful in mechanical system modeling because they can also be used to model the connection of components having relative motion. For example, the device in Fig. 9.5(b), like a spring, has ends that move relative to one another, but the force

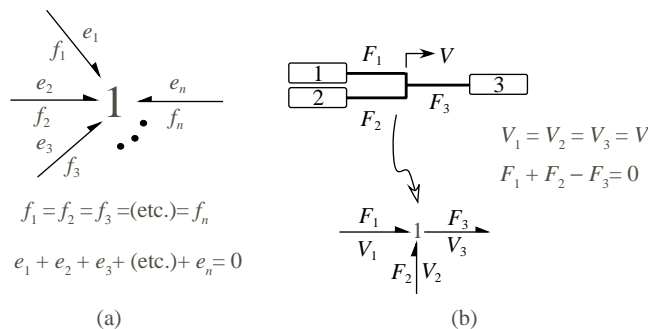



FIGURE 9.4 Mechanical 1-junction: (a) basic definition, (b) example use at a massless junction.

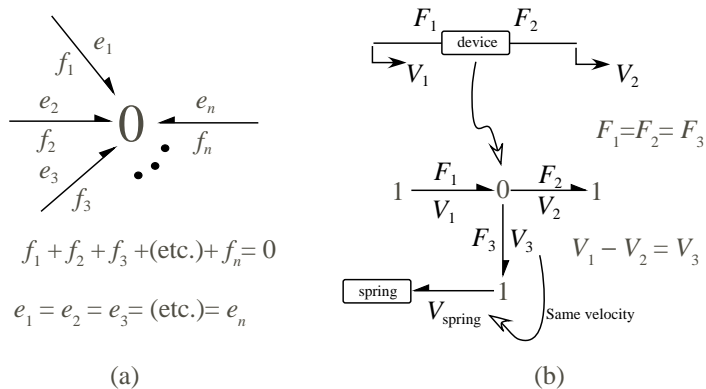



FIGURE 9.5 Mechanical 0-junction: (a) basic definition, (b) example use at a massless junction.

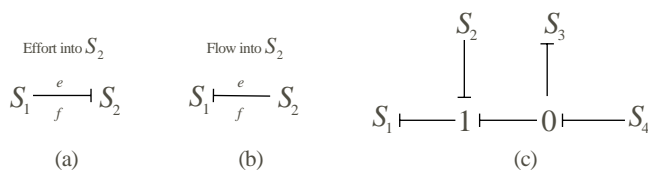



FIGURE 9.6 (a) Specifying effort from S_1 into S_2. (b) Specifying flow from S_1 into S_2. (c) A contrived example showing the constraint on causality assignment imposed by the physical definitions of 0- and 1-junctions.

on each end is the same (note this assumes there is negligible mass). The definition of the 0-junction implies that all the bonds have different velocities, so a flow difference can be formed to construct a relative velocity, V_3. All the bonds have the same force, however, and this force would be applied at the 1-junctions that identify the three distinct velocities in this example. A spring, for example, would be connected on a bond connected to the V_3 junction, as shown in Fig. 9.5(b), and $V_{spring} = V_3$.

The 1- and 0-junction elements graphically represent algebraic structure in a model, with distinct physical attributes from compatibility of kinematics (1-junction) and force or torque (0-junction). The graph should reflect what can be understood about the interconnection of physical devices with a bond graph. There is an advantage in forming a bond graph, since causality can then be used to form mathematical models. See the text by Karnopp, Margolis, and Rosenberg [17] for examples. There is a relation to through and across variables, which are used in linear graph methods [33].

Causality

Bond graph modeling was conceived with a consistent and algorithmic methodology for assignment of causality (see Paynter [28], p. 126). In the context of bond graph modeling, causality refers to the input–output relationship between variables on a power bond, and it depends on the systems connected to each end of a bond. Paynter identified the need for this concept having been extensively involved in analog computing, where solutions rely on well-defined relationships between signals. For example, if system S_1 in Fig. 9.6(a) is a known source of effort, then when connected to a system S_2, it must specify effort into S_2, and S_2 in turn must return the flow variable, f, on the bond that connects the two systems. In a bond graph, this causal relationship is indicated by a vertical stroke drawn on the bond, as shown in Fig. 9.6(a). The vertical stroke at one end of a bond indicates that effort is specified into the multiport element connected at that end. In Fig. 9.6(b), the causality is reversed from that shown in (a).

The example in Fig. 9.6(c) illustrates how causality "propagates" through a bond graph of interconnected bonds and systems. Note that a 1-junction with multiple ports can only have one bond specifying flow at that junction, so the other bonds specify effort into the 1-junction. A 0-junction requires one bond to specify effort, while all others specify flow. Also note that a direction for positive power flow has not been assigned on these bonds. This is intentional to emphasize the fact that power sense and causality assignment on a bond are **independent** of each other.

Causality assignment in system models will be applied in examples that follow. An extensive discussion of the successive cauality assignment procedure (sometimes referred to as SCAP) can be found in Rosenberg and Karnopp [32] or Karnopp, Margolis, and Rosenberg [17]. By using the defined bond graph elements, causality assignment is made systematically. The procedure has been programmed into several commercially available software packages that use bond graphs as formal descriptions of physical system models.

Because it reveals the input–output relationship of variables on all the bonds in a system model, causality can infer computational solvability of a bond graph model. The results are used to indicate the number of dynamic states required in a system, and the causal graph is helpful in actually deriving the mathematical model. Even if equations are not to be derived, causality can be used to derive physical insight into how a system works.

9.3 Descriptions of Basic Mechanical Model Components

Mechanical components in mechatronic systems make their presence known through motional response and by force and torque (or moment) reactions notably on support structures, actuators, and sensors. Understanding and predicting these response attributes, which arise due to combinations of frictional, elastic, and inertial effects, can be gained by identifying their inherent dissipative and energy storing nature. This emphasis on dissipation and energy storage leads to a systematic definition of constitutive relations for basic mechanical system modeling elements. These model elements form the basis for building complex nonlinear system models and for defining impedance relations useful in transfer function formulation. In the following, it is assumed that the system components can be well represented by lumped-parameter formulations.

It is presumed that a modeling decision is made so that dissipative and energy storing (kinetic and potential) elements can be identified to faithfully represent a system of interest. The reticulation is an essential part of the modeling process, but sometimes the definition and interconnection of the elements is not easy or intuitive. This section first reviews mechanical system input and output model elements, and then reviews passive dissipative elements and energy-storing elements. The section also discusses coupling elements used for modeling gears, levers, and other types of power-transforming elements. The chapter concludes by introducing impedance relationships for all of these elements.

Defining Mechanical Input and Output Model Elements

In dynamic system modeling, initial focus requires defining a **system boundary**, a concept borrowed from basic thermodynamics. In isolating mechanical systems, a system boundary identifies ports through which power and signal can pass. Each port is described either by a force–velocity or torque–angular velocity power conjugate pair. It is helpful, when focusing on the mechanical system modeling, to make a judgement on the causality at each port. For example, if a motor is to be attached to one port, it may be possible to define torque as the input variable and angular velocity as the output (back to the motor).

It is important to identify that these are model assumptions. We define specific elements as **sources** of effort or flow that can be attached at the boundary of a system of interest. These inputs might be known and or idealized, or they could simply be "placeholders" where we will later attach a model for an actuator or sensor. In this case, the causality specified at the port is fixed so that the (internal) system model will not change. If the causality changes, it will be necessary to reformulate a new model.

In bond graph terminology, the term **effort source** is used to define an element that specifies an effort, such as this force or torque. The symbol S_e or E can be used to represent the effort source on a bond graph.

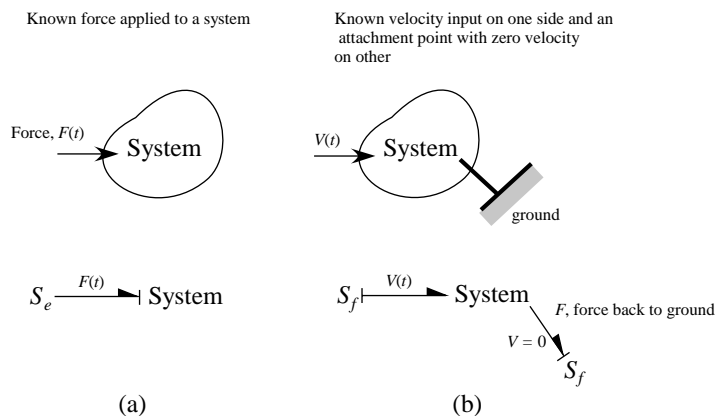



FIGURE 9.7 Two cases showing effort and flow sources on word bond graphs.

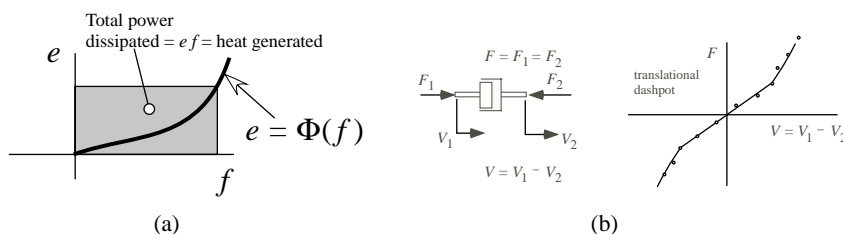



FIGURE 9.8 (a) Resistive constitutive relation. (b) Example dashpot resistive model.

A **flow source** is an element that specifies a flow on a bond, such as a translational velocity or angular or rotational velocity. The bond graph symbol is S_f or F. Two basic examples of sources are shown in Fig. 9.7. Note that each bond has a defined effort or flow, depending on the source type. The causality on these model elements is always known, as shown. Further, each bond carries both pieces of information: (1) the effort or flow variable specified by the source, and (2) the *back reaction* indicated by the causality. So, for example, at the ground connection in Fig. 9.7(b), the source specifies the zero velocity constraint into the system, and the system, in turn, specifies an effort *back* to the ground. The symbolic representation emphasizes the causal nature of bond graph models and emphasizes which variables are available for examination. In this case, the force back into the ground might be a critical output variable.

Dissipative Effects in Mechanical Systems

Mechanical systems will dissipate energy due to friction in sliding contacts, dampers (passive or active), and through interaction with different energy domains (e.g., fluid loading, eddy current damping). These irreversible effects are modeled by constitutive functions between force and velocity or torque and angular velocity. In each case, the product of the effort-flow variables represents power dissipated, $P_d = e \cdot f$, and the total energy dissipated is $E_d = \int P_d \, dt = \int (e \cdot f) \, dt$. This energy can be determined given knowledge of the constitutive function, $e = \Phi(f)$, shown graphically in Fig. 9.8(a). We identify this as a basic *resistive* constitutive relationship that must obey the restriction imposed by the second law of thermodynamics; namely that, $e \cdot f \geq 0$. A typical mechanical dashpot that follows a resistive-type model description is summarized in Fig. 9.8(b).

In a bond graph model, resistive elements are symbolized by an **R** element, and a generalized, multiport **R**-element model is shown in Fig. 9.9(a). Note that the **R** element is distinguished by its ability to represent entropy production in a system. On the **R** element, a *thermal port* and bond are shown, and the power direction is always positive *away* from the **R**. In thermal systems, temperature, T, is the effort variable

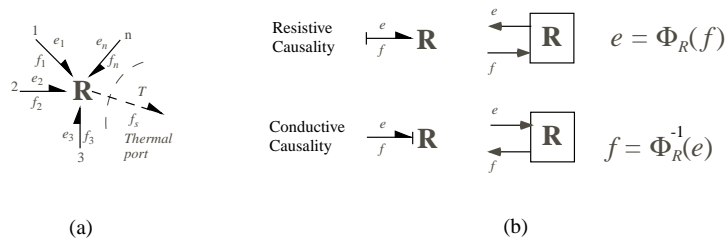



FIGURE 9.9 (a) Resistive bond graph element. (b) Resistive and conductive causality.

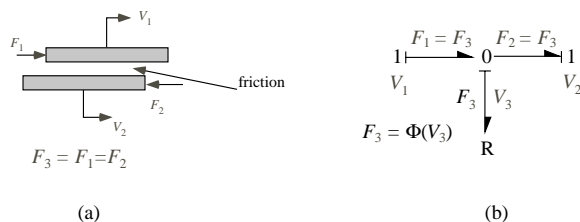
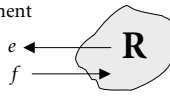
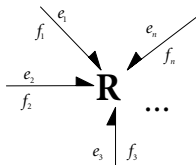
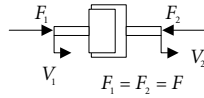
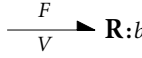
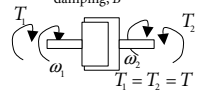
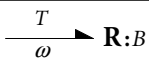



FIGURE 9.10 (a) Two sliding surfaces. (b) Bond graph model with causality implying velocities as known inputs.

and entropy flow rate, f_s is the flow variable. To compute heat generated by the **R** element, compose the calculation as Q (heat in watts) $= T \cdot f_s = \sum_i e_i \cdot f_i$ over the n ports.

The system attached to a resistive element through a power bond will generally determine the causality on that bond, since resistive elements generally have no preferred causal form.[1] Two possible cases on a given **R**-element port are shown in Fig. 9.9(b). A block diagram emphasizes the computational aspect of causality. For example, in a resistive case the flow (e.g., velocity) is a known input, so power dissipated is $P_d = e \cdot f = \Phi(f) \cdot f$. For the linear damper, $F = b \cdot V$, so $P_d = F \cdot V = bV^2$ (W).

In mechanical systems, many frictional effects are driven by relative motion. Hence, identifying how a dissipative effect is configured in a mechanical system requires identifying critical motion variables. Consider the example of two sliding surfaces with distinct velocities identified by 1-junctions, as shown in Fig. 9.10(a). Identifying one surface with velocity V_1, and the other with V_2, the simple construction shown in Fig. 9.10(b) shows how an **R** element can be connected at a *relative* velocity, V_3. Note the relevance of the causality as well. Two velocities join at the 0-junction to form a relative velocity, which is a causal input to the **R**. The causal output is a force, F_3, computed using the constitutive relation, $F = \Phi(V_3)$. The 1-junction formed to represent V_3 can be eliminated when there is only a single element attached as shown. In this case, the **R** would replace the 1-junction.

When the effort-flow relationship is linear, the proportionality constant is a **resistance**, and in mechanical systems these quantities are typically referred to as **damping constants**. Linear damping may arise in cases where two surfaces separated by a fluid slide relative to one another and induce a viscous and strictly laminar flow. In this case, it can be shown that the force and relative velocity are linearly related, and the material and geometric properties of the problem quantify the linear damping constant. Table 9.2 summarizes both translational and rotational damping elements, including the linear cases. These components are referred to as dampers, and the type of damping described here leads to the term viscous friction in mechanical applications, which is useful in many applications involving lubricated surfaces. If the relative speed is relatively high, the flow may become turbulent and this leads to nonlinear damper behavior. The constitutive relation is then a nonlinear function, but the **structure** or interconnection of

[1]This is true in most cases. Energy-storing elements, as will be shown later, have a causal form that facilitates equation formulation.

TABLE 9.2 Mechanical Dissipative Elements

Physical System	Fundamental Relations	Bond Graph
Generalized Dissipative Element (diagram: e, f, **R**) • Resistive element • **Resistance,** R	Dissipation: $\mathbf{e}\cdot\mathbf{f} = \sum_i e_i f_i = T\cdot f_s$ Resistive law: $e = \Phi_R(f)$ Conductive law: $f = \Phi_R^{-1}(e)$ Content: $P_f = \int e \cdot df$ Co-content: $P_e = \int f \cdot de$	(bond graph: e_1, f_1; e_2, f_2; e_3, f_3; e_n, f_n; **R** …) **Generalized multiport R-element**
Mechanical Translation (diagram: damping, b; F_1, F_2, V_1, V_2; $F_1 = F_2 = F$; $V_1 - V_2 = V$) • Damper • **damping,** b	Constitutive: $F = \Phi(V)$ Content: $P_V = \int F \cdot dV$ Co-energy: $P_F = \int V \cdot dF$ Dissipation: $P_d = P_V + P_F$	(bond: F, V → **R**:b) Linear: $F = b\cdot V$ Dissipation: $P_d = bV^2$
Mechanical Rotation (diagram: damping, B; T_1, T_2, ω_1, ω_2; $T_1 = T_2 = T$; $\omega_1 - \omega_2 = \omega$) • Torsional damper • **damping,** B	Constitutive: $T = \Phi(\omega)$ Content: $P_\omega = \int T \cdot d\omega$ Co-energy: $P_T = \int \omega \cdot dT$ Dissipation: $P_d = P_\omega + P_T$	(bond: T, ω → **R**:B) Linear: $T = B\cdot\omega$ Dissipation: $P_d = B\omega^2$

TABLE 9.3 Typical Coefficient of Friction Values. Note, Actual Values Will Vary Significantly Depending on Conditions

Contacting Surfaces	Static, μ_s	Sliding or Kinetic, μ_k
Steel on steel (dry)	0.6	0.4
Steel on steel (greasy)	0.1	0.05
Teflon on steel	0.04	0.04
Teflon on teflon	0.04	—
Brass on steel (dry)	0.5	0.4
Brake lining on cast iron	0.4	0.3
Rubber on asphalt	—	0.5
Rubber on concrete	—	0.6
Rubber tires on smooth pavement (dry)	0.9	0.8
Wire rope on iron pulley (dry)	0.2	0.15
Hemp rope on metal	0.3	0.2
Metal on ice	—	0.02

the model in the system does not change. Dampers are also constructed using a piston/fluid design and are common in shock absorbers, for example. In those cases, the force–velocity characteristics are often tailored to be nonlinear.

The viscous model will not effectively model friction between dry solid bodies, which is a much more complex process and leads to performance bounds especially at lower relative velocities. One way to capture this type of friction is with the classic Coulomb model, which depends on the normal load between surfaces and on a coefficient of friction, typically denoted μ (see Table 9.3). The Coulomb model quantifies the friction force as $F = \mu N$, where N is the normal force. This function is plotted in Fig. 9.11(a) to illustrate how it models the way the friction force always opposes motion. This model still qualifies as a resistive constitutive function relating the friction force and a relative velocity of the surfaces. In this case,

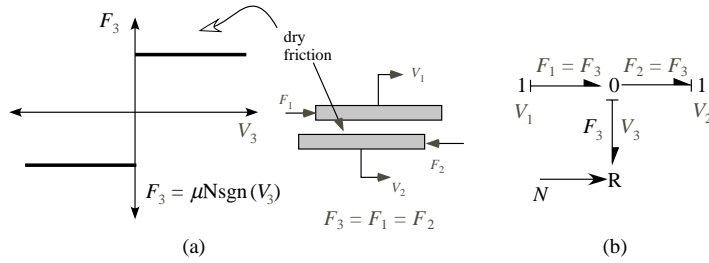



FIGURE 9.11 (a) Classic coulomb friction for sliding surfaces. (b) Bond graph showing effect of normal force as a modulation of the **R**-element law.

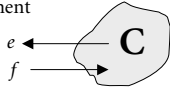
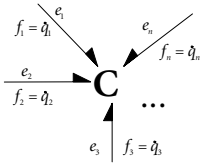
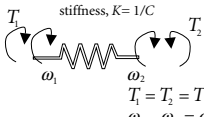
however, the velocity comes into effect only to determine the sign of the force; i.e., $F = \mu N \,\text{sgn}(V)$, where sgn is the signum function (value of 1 if $V > 0$ and -1 if $V < 0$).

This model requires a special condition when $V \rightarrow 0$. Dry friction can lead to a phenomenon referred to as stick-slip, particularly common when relative velocities between contacting surfaces approach low values. Stick-slip, or stiction, friction forces are distinguished by the way they vary as a result of other (modulating) variables, such as the normal force or other applied loads. Stick-slip is a type of system response that arises due to frictional effects. On a bond graph, a signal bond can be used to show that the normal force is determined by an external factor (e.g., weight, applied load, etc.). This is illustrated in Fig. 9.11(b). When the basic properties of a physical element are changed by signal bonds in this way, they are said to be **modulated**. This is a modeling technique that is very useful, but care should be taken so it is not applied in a way that violates basic energy principles.

Another difficulty with the standard dry friction model is that it has a preferred causality. In other words, if the causal input is velocity, then the constitutive relation computes a force. However, if the causal input is force then there is no unique velocity output. The function is not bi-unique. Difficulties of this sort usually indicate that additional underlying physical effects are not modeled. While the effort-flow constitutive relation is used, the form of the constitutive relation may need to be parameterized by other critical variables (temperature, humidity, etc.). More detailed models are beyond the scope of this chapter, but the reader is referred to Rabinowicz (1995) and Armstrong-Helouvry (1991) who present thorough discussions on modeling friction and its effects. Friction is usually a dominant source of uncertainty in many predictive modeling efforts (as is true in most energy domains).

Potential Energy Storage Elements

Part of the energy that goes into deforming any mechanical component can be associated with pure (lossless) storage of potential energy. Often the decision to model a mechanical component this way is identified through a basic constitutive relationship between an effort variable, e (force, torque), and a displacement variable, q (translational displacement, angular displacement). Such a relationship may be derived either from basic mechanics [29] or through direct measurement. An example is a translational spring in which a displacement of the ends, x, is related to an applied force, F, as $F = F(x)$.

In an energy-based lumped-parameter model, the generalized displacement variable, q, is used to define a state-determined potential energy function,

$$E = E(q) = U_q$$

This energy is related to the constitutive relationship, $e = \mathsf{F}(q)$, by

$$U(q) = {}_qU = \int e\,dq \;=\; \int \Phi(q)\,d$$

It is helpful to generalize in this way, and to identify that practical devices of interest will have at least one connection (or port) in which power can flow to store potential energy. At this port the displacement

TABLE 9.4 Mechanical Potential Energy Storage Elements (Integral Form)

Physical System	Fundamental Relations	Bond Graph
Generalized Potential Energy Storage Element e, f, **C** • Capacitive element • **Capacitance,** C	State: $\mathbf{q} = \text{displacement}$ Rate: $\dot{\mathbf{q}} = \mathbf{f}$ Constitutive: $\mathbf{e} = \Phi(\mathbf{q})$ Energy: $U_{\mathbf{q}} = \int \mathbf{e} \cdot d\mathbf{q}$ Co-energy: $U_{\mathbf{e}} = \int \mathbf{q} \cdot d\mathbf{e}$	e_1, $f_1 = \dot{q}_1$; e_2, $f_2 = \dot{q}_2$; e_3, $f_3 = \dot{q}_3$; e_n, $f_n = \dot{q}_n$; **C** … **Generalized multiport C-element**
Mechanical Translation stiffness, $k = 1/C$ F_1, F_2, V_1, V_2 $F_1 = F_2 = F$ • spring $V_1 - V_2 = V$ • **stiffness,** k**, compliance,** C	State: $x = \text{displacement}$ Rate: $\dot{x} = V$ Constitutive: $F = F(x)$ Energy: $U_x = \int F \cdot dx$ Co-energy: $U_F = \int x \cdot dF$	F, $\dot{x} = V$, **C**$:1/C = k$ Linear: $F = k \cdot x$ Energy: $U_x = \frac{1}{2}kx^2$ Co-energy: $U_F = F^2/2k$
Mechanical Rotation stiffness, $K = 1/C$ T_1, T_2, ω_1, ω_2 $T_1 = T_2 = T$ $\omega_1 - \omega_2 = \omega$ • Torsional spring • **stiffness,** K**, compliance,** C	State: $\theta = \text{angle}$ Rate: $\dot{\theta} = \omega$ Constitutive: $T = T(\theta)$ Energy: $U_\theta = \int T \cdot d\theta$ Co-energy: $U_T = \int \theta \cdot dT$	T, $\dot{\theta} = \omega$, **C**$:1/C = K$ Linear: $T = K \cdot \theta$ Energy: $U_\theta = \frac{1}{2}k\theta^2$ Co-energy: $U_T = T^2/2K$

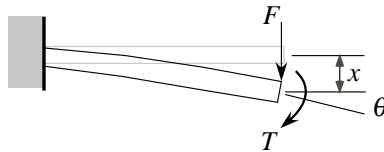
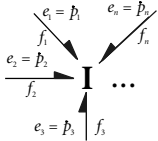
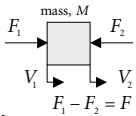
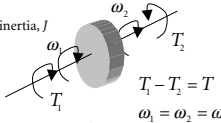



FIGURE 9.12 Example of two-port potential energy storing element: (a) cantilevered beam with translational and rotational end connections, (b) **C**-element, 2-port model.

variable of interest is either translational, x, or angular, θ, and the associated velocities are $V = \dot{x}$ and $\omega = \theta$, respectively. A generalized potential energy storage element is summarized in Table 9.4, where examples are given for the translational and rotational one-port.

The linear translational spring is one in which $F = F(x) = kx = (1/C)x$, where k is the stiffness and $C \equiv 1/k$ is the *compliance* of the spring (compliance is a measure of "softness"). As shown in Table 9.4, the potential energy stored in a linear spring is $U_x = \int F\,dx = \int kx\,dx = \frac{1}{2}kx^2$, and the co-energy is $U_F = \int F\,dx = \int (F/k)\,dF = F^2/2k$. Since the spring is linear, you can show that $U_x = U_F$. If the spring is nonlinear due to, say, plastic deformation or work hardening, then this would not be true.

Elastic potential energy can be stored in a device through multiple ports and through different energy domains. A good example of this is the simple cantilevered beam having both tip force and moment (torque) inputs. The beam can store energy either by translational or rotational displacement of the tip. A constitutive relation for this 2-port **C**-element relates the force and torque to the linear and rotational displacments, as shown in Fig. 9.12. A stiffness (or compliance) matrix for small deflections is derived by linear superposition.

Kinetic Energy Storage

All components that constitute mechanical systems have mass, but in a system analysis, where the concern is dynamic performance, it is often sufficient to focus only on those components that may store relevant amounts of kinetic energy through their motion. This presumes that an energetic basis is used for modeling, and that the tracking of kinetic energy will provide insight into the system dynamics. This is the focus of this discussion, which is concerned for the moment with one-dimensional translation and fixed-axis rotation. Later it will be shown how the formulation presented here is helpful for understanding more complex systems.

The concept of mass and its use as a model element is faciliated by Newton's relationship between the rate of change of momentum of the mass to the net forces exerted on it, $F = \dot{p}$, where p is the momentum. The energy stored in a system due to translational motion with velocity V is the kinetic energy. Using the relation from Newton's law, $dp = Fdt$, this energy is $E(p) = T(p) = T_p = \int P dt = \int FV \ dt = \int V \ dp$. If the velocity is expressed solely as a function of the momentum, p, this system is a pure translational mass, $V = \Phi(p)$. If the velocity is linearly proportional to the momentum, then $V = p/m$, where m is the mass. Similar basic definitions are made for a body in rotation about a fixed axis, and these elements are summarized in Table 9.5.

For many applications of practical interest to engineering, the velocity–momentum relation, $V = V(p)$ (the constitutive relation), is linear. Only in relativistic cases might there be a nonlinear relationship in the constitutive law for a mass. Nevertheless, this points out that for the general case of kinetic energy storage a constitutive relation is formed between the flow variable and the momentum variable, $f = f(p)$. This should help build appreciation for analogies with other energy domains, particularly in electrical systems where inductors (the mass analog) can have nonlinear relationships between current (a flow) and flux linkage (momentum).

The rotational motion of a rigid body considered here is constrained thus far to the simple case of planar and fixed-axis rotation. The mass moment of intertia of a body about an axis is defined as the sum of the products of the mass-elements and the squares of their distance from the axis. For the discrete case, $I = \sum r^2 \Delta m$, which for continuous cases becomes, $I = \int r^2 dm$ (units of kg m^2). Some common shapes

TABLE 9.5 Mechanical Kinetic Energy Storage Elements (Integral Form)

Physical System	Fundamental Relations	Bond Graph
Generalized Kinetic Energy Storage Element (e, f, **I**) • Inertive element • **Inertance, I**	State: $\mathbf{p}$ = momentum Rate: $\dot{\mathbf{p}} = \mathbf{e}$ Constitutive: $\mathbf{f} = \Phi(\mathbf{p})$ Energy: $T_p = \int \mathbf{f} \cdot d\mathbf{p}$ Co-energy: $T_f = \int \mathbf{p} \cdot d\mathbf{f}$	$e_1 = \dot{p}_1$, f_1; $e_n = \dot{p}_n$, f_n; $e_2 = \dot{p}_2$, f_2; $e_3 = \dot{p}_3$, f_3; **I** ... **Generalized multiport I-element**
Mechanical Translation (mass, M; F_1, F_2; V_1, V_2) $F_1 - F_2 = F$ $V_1 = V_2 = V$ • Mass • **mass, m**	State: p = momentum Rate: $\dot{p} = F$ Constitutive: $V = V(p)$ Energy: $T_p = \int f \cdot dp$ Co-energy: $T_V = \int p \cdot dV$	$\dot{p} = F$, V → **I**:M Linear: $V = p/M$ Energy: $T_p = p^2/2M$ Co-energy: $T_V = \frac{1}{2} M V^2$
Mechanical Rotation (inertia, J; ω_1, ω_2; T_1, T_2) $T_1 - T_2 = T$ $\omega_1 = \omega_2 = \omega$ • Rotational inertia • **mass moment of inertia, J**	State: h = angular momentum Rate: $\dot{h} = T$ Constitutive: $\omega = \omega(h)$ Energy: $T_h = \int \omega \cdot dh$ Co-energy: $T_\omega = \int h \cdot d\omega$	$\dot{h} = T$, ω → **I**:J Linear: $\omega = h/J$ Energy: $T_h = h^2/2J$ Co-energy: $T_\omega = \frac{1}{2} J \omega^2$

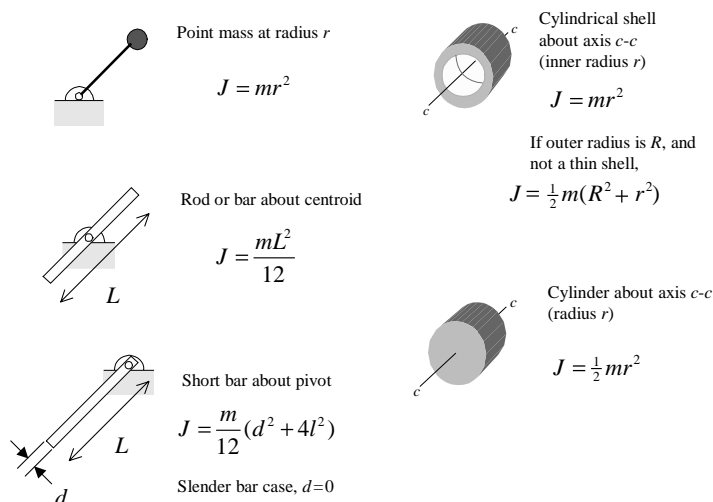



FIGURE 9.13 Mass moments of inertia for some common bodies.

and associated mass moments of inertia are given in Fig. 9.13. General rigid bodies are discussed in section "Inertia Properties."

There are several useful concepts and theorems related to the properties of rigid bodies that can be helpful at this point. First, if the mass moment of inertia is known about an axis through its center of mass (I_G), then Steiner's theorem (parallel axis theorem) relates this moment of inertia to that about another axis a distance d away by $I = I_G + md^2$, where m is the mass of the body. It is also possible to build a moment of inertia for composite bodies, in those situations where the individual motion of each body is negligible. A useful concept is the radius of gyration, k, which is the radius of an imaginary cylinder of infinitely small wall thickness having the same mass, m, and the same mass moment of inertia, I, as a body in question, and given by, $k = \sqrt{I/m}$. The radius of gyration can be used to find an equivalent mass for a rolling body, say, using $m_{eq} = I/k^2$.

Coupling Mechanisms

Numerous types of devices serve as couplers or power transforming mechanisms, with the most common being levers, gear trains, scotch yokes, block and tackle, and chain hoists. Ideally, these devices and their analogs in other energy domains are power conserving, and it is useful to represent them using a 2-port model. In such a model element, the power in is equal to the power out, or in terms of effort-flow pairs, $e_1 f_1 = e_2 f_2$. It turns out that there are two types of basic devices that can be represented this way, based on the relationship between the power variables on the two ports. For either type, a relationship between two of the variables can usually be identified from geometry or from basic physics of the device. By imposing the restriction that there is an ideal power-conserving transformation inherent in the device, a second relationship is derived. Once one relation is established the device can usually be classified as a **transformer** or **gyrator**. It is emphasized that these model elements are used to represent the ideal power-conserving aspects of a device. Losses or dynamic effects are added to model real devices.

A device can be modeled as a **transformer** when $e_1 = me_2$ and $mf_1 = f_2$. In this relation, m is a transformer **modulus** defined by the device physics to be constant or in some cases a function of states of the system. For example, in a simple gear train the angular velocities can be ideally related by the ratio of pitch radii, and in a slider crank there can be formed a relation between the slider motion and the crank angle. Consequently, the two torques can be related, so the gear train is a transformer. A device can be modeled as a **gyrator** if $e_1 = rf_2$ and $rf_1 = e_2$, where r is the gyrator modulus. Note that this model can represent

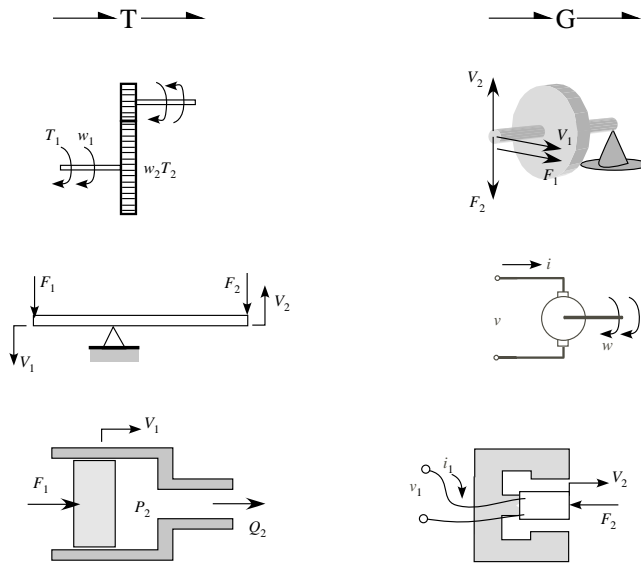



FIGURE 9.14 Common devices that can be modeled as transformers and gyrators in mechatronic systems.

the power-conserving transformation in devices for which a cross-relationship between power variables (i.e., effort related to flow) has been identified.[2]

Some examples of transformers and gyrators are shown in Fig. 9.14. In a bond graph model, the transformer can be represented by a **TF** or **T**, while a gyrator is represented by a **GY** or **G** (note, the two letter symbol is common). The devices shown in Fig. 9.14 indicate a modulus *m* or *r*, which may or may not be a constant value. Many devices may have power-conserving attributes; however, the relationship between the effort-flow variables may not be constant, so the relationship is said to be *modulated* when the modulus is a function of a dynamic variable (preferably a state of the system). On a bond graph, this can be indicated using a signal bond directed into the **T** or **G** modulus.

Examples of a modulated transformer and gyrator are given in Fig. 9.15. These examples highlight useful techniques in modeling of practical devices. In the slider crank, note that the modulation is due to a change in the angular position of the crank. We can get this information from a bond that is adjacent to the transformer in question; that is, if we integrate the angular velocity found on a neighboring bond, as shown in Fig. 9.15(a). For the field excited dc motor shown in Fig. 9.15(b), the torque–current relation in the motor depends on a flux generated by the field; however, this field is excited by a circuit that is powered *independent* of the armature circuit. The signal information for modulation does not come from a neighboring bond, as in the case for the slider crank. These two examples illustrate two ways that constraints are imposed in coupling mechanisms. The modulation in the slider crank might be said to represent a holonomic constraint, and along these same lines the field excitation in the motor imposes a non-holonomic constraint. We cannot relate torque and current in the latter case without solving for the dynamics of an independent system—the field circuit. In the slider crank, the angular position required for the modulation is obtained simply by integrating the velocity, since $\dot{\theta} = \omega$. Additional discussion on constraints can be found in section 9.7.

The system shown in Fig. 9.16(a) is part of an all-mechanical constant-speed drive. A mechanical feedback force, F_2, will adjust the position of the middle rotor, x_2. The effect is seen in the bond graph

[2]It turns out that the gyrator model element is essential in all types of systems. The need for such an element to represent gyroscopic effects in mechanical systems was first recognized by Thomson and Tait in the late 1900s. However, it was G. D. Birkhoff (1927) and B. D. H. Tellegen (1948) who independently identified the need for this element in analysis and synthesis of systems.

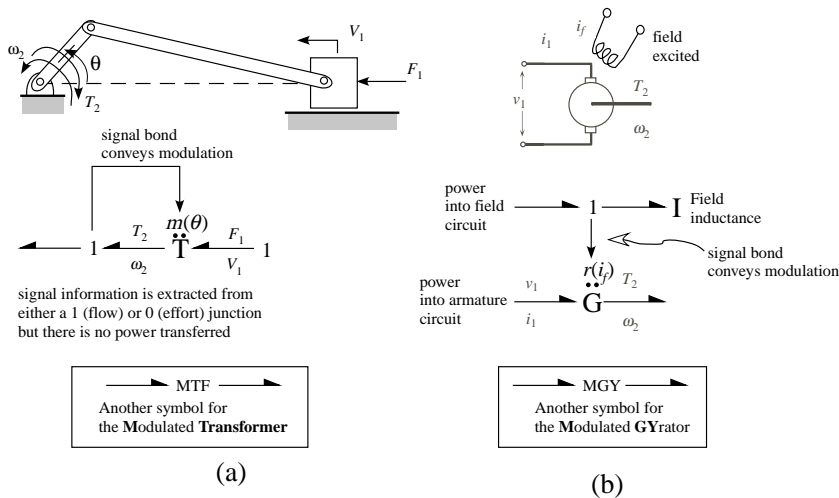



FIGURE 9.15 Concept of modulation in transformers and gyrators.

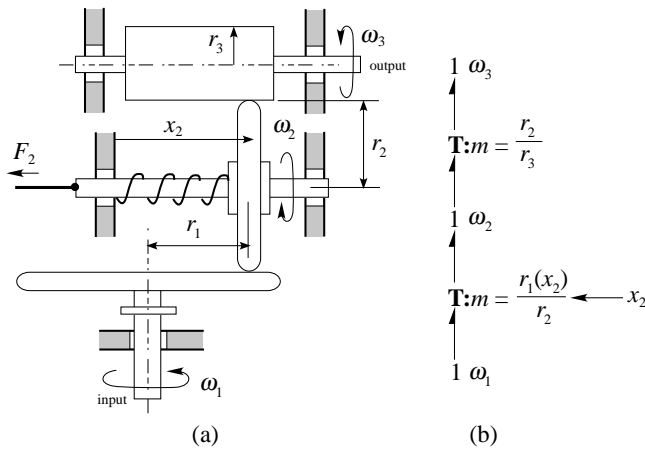



FIGURE 9.16 A nonholonomic constraint in a transformer model.

model of Fig. 9.16(b), which has two transformers to represent the speed ratio between the input (turntable) 1 and the mid-rotor 2, and the speed ratio between the mid-rotor and the output roller 3. The first transformer is a mechanical version of a nonholonomic transformation. Specifically, we would have to solve for the dynamics of the rotor position (x_2) in order to transform power between the input and output components of this device.

Impedance Relationships

The basic component descriptions presented so far are the basis for building basic models, and a very useful approach relies on impedance formulations. An impedance function, Z, is a ratio of effort to flow variables at a given system port of a physical device, and the most common application is for linear systems where $Z = Z(s)$, where s is the complex frequency variable (sometimes called the Laplace operator). An admittance is the inverse of the impedance, or $Y = 1/Z$. For each basic element defined, a linear impedance relation can be derived for use in model development. First, recall that the derivative operator can be represented by the s operator, so that dx/dt in s-domain is simply sx and $\int x\,dt$ is x/s, and so on.

TABLE 9.6 Basic Mechanical Impedance Elements

System	Resistive, Z_R	Capacitive, Z_C	Inertive, Z_I
Translation	b	k/s	$m \cdot s$
Rotation	B	K/s	$J \cdot s$

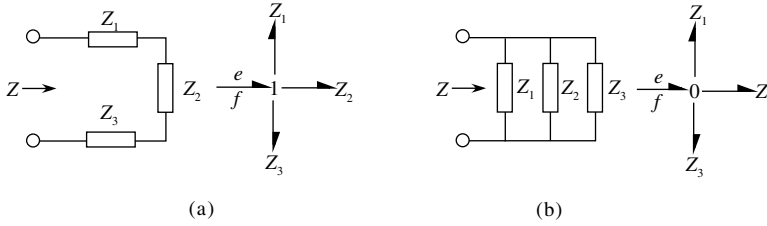



FIGURE 9.17 (a) Impedance of a series connection. (b) Admittance for a parallel combination.

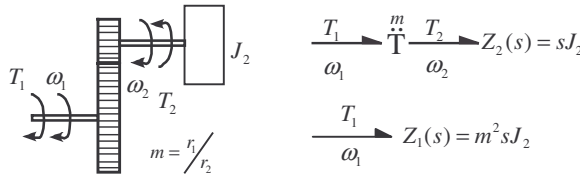



FIGURE 9.18 Rotational inertia attached to gear train, and corresponding model in impedance form. This example illustrates how a transformer can scale the gain of an impedance.

For the basic inertia element in rotation, for example, the basic rate law (see Table 9.5) is $\dot{h} = T$. In s-domain, $sh = T$. Using the linear constitutive relation, $h = J\omega$, so $sJ\omega = T$. We can observe that a rotation inertial impedance is defined by taking the ratio of effort to flow, or $T/\omega \equiv Z_I = sJ$. A similar exercise can be conducted for every basic element to construct Table 9.6.

Using the basic concept of a 0 junction and a 1 junction, which are the analogs of parallel and series circuit connections, respectively, basic impedance formulations can be derived for bond graphs in a way analogous to that done for circuits. Specifically, when impedances are connected in series, the total impedance is the sum, while admittances connected in parallel sum to give a total admittance. These basic relations are illustrated in Fig. 9.17, for which

$$Z = \underbrace{Z_1 + Z_2 + \cdots + Z_n}_{\substack{n \text{ impedances in series sum} \\ \text{to form a total impedance}}}, \qquad Y = \underbrace{Y_1 + Y_2 + \cdots + Y_n}_{\substack{n \text{ admittances in parallel sum} \\ \text{to form a total admittance}}} \tag{9.2}$$

Impedance relations are useful when constructing transfer functions of a system, as these can be developed directly from a circuit analog or bond graph. The transformer and gyrator elements can also be introduced in these models. A device that can be modeled with a transformer and gyrator will exhibit impedance-scaling capabilities, with the moduli serving a principal role in adjusting how an impedance attached to one "side" of the device appears when "viewed" from the other side. For example, for a device having an impedance Z_2 attached on port 2, the impedance as viewed from port 1 is derived as

$$Z_1 = \frac{e_1}{f_1} = \left[\frac{e_1}{e_2}\right]\left[\frac{e_2}{f_2}\right]\left[\frac{f_2}{f_1}\right] = [m][Z_2(s)][m] = m^2 Z_2(s) \tag{9.3}$$

This concept is illustrated by the gear-train system in Fig. 9.18. A rotational inertia is attached to the output shaft of the gear pair, which can be modeled as a transformer (losses, and other factors ignored here).

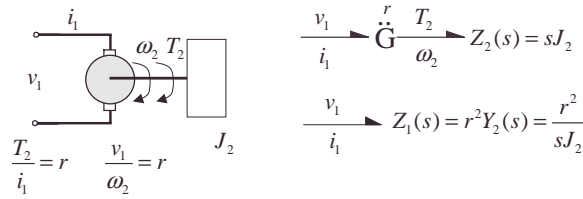



FIGURE 9.19 Rotational inertial attached to a basic rotational machine modeled as a simple gyrator. This example illustrates how a gyrator can scale the gain but also convert the impedance to an admittance form.

The impedance of the inertial is $Z_2 = sJ_2$, where J_2 is the mass moment of inertia. The gear train has an impedance-scaling capability, which can be designed through selection of the gear ratio, m.

The impedance change possible with a transformer is only in gain. The gyrator can affect gain and in addition can change the impedance into an admittance. Recall the basic gyrator relation, $e_1 = rf_2$ and $e_2 = rf_1$, then for a similar case as before,

$$\frac{1}{1} = \left[\frac{e_1}{f_2}\right]\left[\frac{f_2}{e_2}\right]\left[\frac{e_2}{f_1}\right] = [r][Y_2(s)][r] = r^2 Y_2(s \tag{9.4}$$

This functional capability of gyrators helps identify basic motor-generator designs as integral parts of a flywheel battery system. A very simplified demonstration is shown in Fig. 9.19, where a flywheel (rotational inertia) is attached to the mechanical port of a basic electromechanical gyrator. When viewed from the electrical port, you can see that the gyrator makes the inertia "look" like a potential energy storing device, since the impedance goes as $1/(sC)$, like a capacitive element, although here C is a mechanical inertia.

9.4 Physical Laws for Model Formulation

This section will illustrate basic equation formulation for systems ranging in complexity from mass-spring-damper models to slightly more complex models, showing how to interface with nonmechanical models.

Previous sections of this chapter provide descriptions of basic elements useful in modeling mechanical systems, with an emphasis on a dynamic system approach. The power and energy basis of a bond graph approach makes these formulations consistent with models of systems from other energy domains. An additional benefit of using a bond graph approach is that a systematic method for causality assignment is available. Together with the physical laws, causal assignment provides insight into how to develop computational models. Even without formulating equations, causality turns out to be a useful tool.

Kinematic and Dynamic Laws

The use of basic kinematic and dynamic equations imposes a structure on the models we build to represent mechanical translation and rotation. Dynamic equations are derived from Newton's laws, and we build free-body diagrams to understand how forces are imposed on mechanical systems. In addition, we must use geometric aspects of a system to develop kinematic equations, relying on properly defined coordinate systems. If the goal is to analyze a mechanical system alone, typically the classical application of conservation of momentum or energy methods and/or the use of kinematic analysis is required to arrive at solutions to a given problem. In a mechatronic system, it is implied that a mechanical system is coupled to other types of systems (hydraulics, electromechanical devices, etc.). Hence, we focus here on how to build models that will be easily integrated into overall system models. A detailed classical discussion of kinematics and dynamics from a fundamental perspective can be found in many introductory texts such as Meriam and Kraige [23] and Bedford and Fowler [5], or in more advanced treatments by Goldstein [11] and Greenwood [12].

When modeling simple translational systems or fixed-axis rotational systems, the basic set of laws summarized below are sufficient to build the necessary mathematical models.

Basic Dynamic and Kinematic Laws

System	Dynamics	Kinematics
Translational	$\sum_i^N F_i = 0$	$\sum_i^N V_i = 0$
Rotational	$\sum_i^N T_i = 0$	$\sum_i^N \omega_i = 0$
Junction type	1-junction	0-junction

There is a large class of mechanical systems that can be represented using these basic equations, and in this form it is possible to see how: (a) bond graph junction elements can be used to structure these models and (b) how these equations support circuit analog equations, since they are very similar to the Kirchhoff circuit laws for voltage and current. We present here the bond graph approach, which graphically communicates these physical laws through the 0- and 1-junction elements.

Identifying and Representing Motion in a Bond Graph

It is helpful when studying a mechanical system to focus on identifying points in the system that have distinct velocities (V or ω). One simply can associate a 1-junction with these points. Once this is done, it becomes easier to identify connection points for other mechanical components (masses, springs, dampers, etc.) as well as points for attaching actuators or sensors. Further, it is critical to identify and to define additional velocities associated with relative motion. These may not have clear, physically identifiable points in a system, but it is necessary to localize these in order to attach components that rely on relative motion to describe their operation (e.g., suspensions).

Figure 9.20 shows how identifying velocities of interest can help identify 1-junctions at which mechanical components can be attached. For the basic mass element in part (a), the underlying premise is that a component of a system under study is idealized as a pure translational mass for which momentum and velocity are related through a constitutive relation. What this implies is that the velocity of the mass is the same throughout this element, so a 1-junction is used to identify this distinct motion. A bond attached to this 1-junction represents how any power flowing into this junction can flow into a kinetic energy storing element, **I**, which represents the mass, m. Note that the force on the bond is equal to the rate of change of momentum, $\dot{p}$, where $p = mV$.

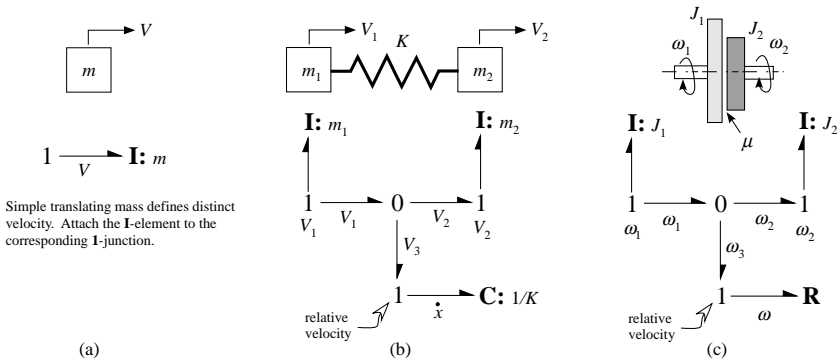
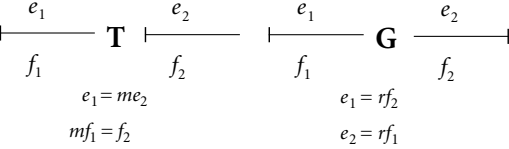
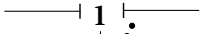
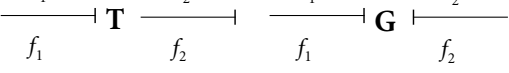



FIGURE 9.20 Identifying velocities in a mechanical system can help identify correct interconnection of components and devices: (a) basic translating mass, (b) basic two-degree of freedom system, (c) rotational frictional coupling between two rotational inertias.

The two examples in Figs. 9.20(b) and 9.20(c) demonstrate how a relative velocity can be formed. Two masses each identify the two distinct velocity points in these systems. Using a 0-junction allows construction of a *velocity difference*, and in each case this forms a relative velocity. In each case the relative velocity is represented by a 1-junction, and it is critical to identify that this 1-junction is essentially an attachment point for a basic mechanical modeling element.

Assigning and Using Causality

Bond graphs describe how modeling decisions have been made, and how model elements (**R**, **C**, etc.) are interconnected. A power bond represents power flow, and assigning power convention using a half-arrow is an essential part of making the graph useful for modeling. A sign convention is essential for expressing the algebraic summation of effort and flow variables at 0- and 1-junctions. Power is generally assigned positive sense flowing into passive elements (resistive, capacitive, inertive), and it is usually safe to always adopt this convention. Sign convention requires consistent and careful consideration of the reference conditions, and sometimes there may be some arbitrariness, not unlike the definition of reference directions in a free-body diagram.

Causality involves an augmentation of the bond graph, but is strictly independent of power flow convention. As discussed earlier, an assignment is made on each bond that indicates the input–output relationship of the effort-flow variables. The assignment of causality follows a very consistent set of rules. A system model that has been successfully assigned causality on all bonds essentially communicates solvability of the underlying mathematical equations. To understand where this comes from, we can begin by examining the contents of Tables 9.4 and 9.5. These tables refer to the *integral* form of the energy storage elements. An energy storage element is in integral form if it has been assigned integral causality. Integral causality implies that the causal input variable (effort or flow) leads to a condition in which the state of the energy stored in that element can be determined only by *integrating* the fundamental rate law. As shown in Table 9.7, integral causality for an **I** element implies effort is the input, whereas integral causality for the **C** element implies flow is the input.

TABLE 9.7 Table Summarizing Causality for Energy Storage Elements

TABLE 9.8 Table of Causality Assignment Guidelines

FIGURE 9.21 Driving a rotational inertia with a velocity source: (b) simple bond graph with causality, (c) explanation of back effect.

As shown in this table, the alternative causality for each element leads to *derivative causality*, a condition in which the state of the energy storage element is known instantaneously and as such is said to be *dependent* on the input variable, and is in a state of dependent causality. The implication is that energy storage elements in integral causality require one differential equation (the rate law) to be solved in order to determine the value of the *state variable* (p or q). Energy storage elements in derivative causality don't require a differential equation; however, they still make their presence known through the back reaction implied. For example, if an electric machine shown in Fig. 9.21(a) is assumed to drive a rotational inertial with a known velocity, ω, then the inertia is in derivative causality. There will also be losses, but the problem is simplified to demonstrate the causal implications. The energy is always known since, $h = J\omega$, so $T_h = h^2/2J$. However, the machine will feel an inertial back torque, $\dot{h}$, whenever a change is made to ω. This effect cannot be neglected.

Causality assignment on some of the other modeling elements is very specific, as shown in Table 9.8. For example, for sources of effort or flow, the causality is implied. On the two-port transformer and gyrator, there are two possible causality arrangements for each. Finally, for 0- and 1-junctions, the causality is also very specific since in each case only one bond can specify the effort or flow at each.

With all the guidelines established, a basic causality assignment procedure can be followed that will make sure all bonds are assigned causality (see also Rosenberg and Karnopp [32] and Karnopp, Margolis, and Rosenberg [17]).

1. For a given system, assign causality to any effort or flow sources, and for each one assign the causality as required through 0- and 1-junctions and transformer and gyrator elements. The causality should be spread through the model until a point is reached where no assignment is implied. Repeat this procedure until all sources have been assigned causality.
2. Assign causality to any **C** or **I** element, trying to assign integral causality if possible. For each assignment, propagate the causality through the system as required. Repeat this procedure until all storage elements are assigned causality.

3. Make any final assignments on **R** elements that have not had their causality assigned through steps 1 and 2, and again propagate causality as required. Any arbitrary assignment on an **R** element will indicate need for solving an algebraic equation.
4. Assign any remaining bonds arbitrarily, propagating each case as necessary.

Causality can provide information about system operation. In this sense, the bond graph provides a picture of how inputs to a system lead to certain outputs. The use of causality with a bond graph replaces ad hoc assignment of causal notions in a system. This type of information is also useful for understanding how a system can be split up into modules for simulation and/or it can confirm the actual physical boundaries of components.

Completing the assignment of causality on a bond graph will also reveal information about the solvability of the system model. The following are key results from causality assignment.

- Causality assignment will reveal the order of the system, which is equal to the number of independent energy storage elements (i.e., those with integral causality). The state variable (p or q) for any such element will be a state of the system, and one first-order differential equation will be required to describe how this state propagates through time.
- Any arbitrary assignment of causality on an **R** element indicates there is an algebraic loop. The number of arbitrary assignments can be related to the number of algebraic equations required in the model.

Developing a Mathematical Model

Mathematical models for lumped-parameter mechanical systems will take the form of coupled ordinary differential equations or, for a linear or linearized system, transfer functions between variables of interest and system inputs. The form of the mathematical model should match the application, and one can readily convert between the different forms. A classical approach to developing the mathematical model will involve applying Newton's second law directly to each body, taking account of the forces and torques. Commonly, the result is a second-order ordinary differential equation for each body in a system. An alternative is to use Lagrange's equations, and for multidimensional dynamics, where bodies may have combined translation and rotation, additional considerations are required as will be discussed in Section 9.6. At this point, consider those systems where a given body is either under translation or rotation.

Mass-Spring-Damper: Classical Approach

A basic mechanical system that consists of a rigid body that can translate in the z-direction is shown in Fig. 9.22(a). The system is modeled using a mass, a spring, and a damper, and a force, $F(t)$, is applied

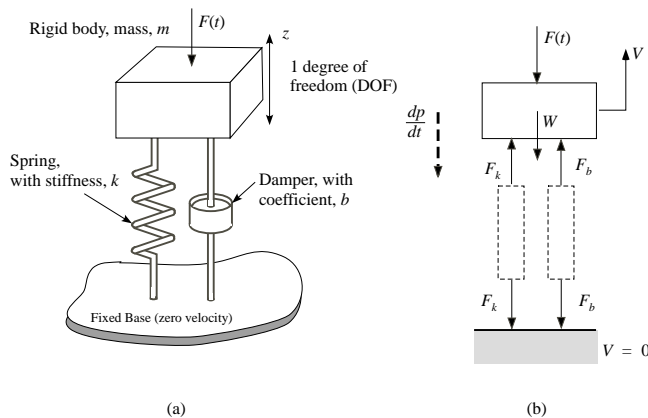



FIGURE 9.22 Basic mass-spring-damper system: (a) schematic, (b) free-body diagram.

directly to the mass. A free-body diagram in part (b) shows the forces exerted on the system. The spring and damper exert forces F_k and F_b on the mass, and these same forces are also exerted on the fixed base since the spring and damper are assumed to be massless. A component of the weight, W, resolved along the axis of motion is included. The sum of applied forces is then, $\sum F = F(t) + W - F_k - F_b$. The dashed arrow indicates the "inertial force" which is equal to the rate of change of the momentum in the z-direction, p_z, or, $dp_z/dt = \dot{p}_z = m\dot{V}_z$. This term is commonly used in a D'Alembert formulation, one can think of this force as opposing or resisting the effect of applied forces to accelerate the body. It is common to use the inertial force as an "applied force," especially when performing basic analysis (e.g., see Chapter 3 or 6 of [23]).

Newton's second law relates rate of change of momentum to applied forces, $\dot{p} = \sum F$, so, $\dot{p}_z = F(t) + W - F_k - F_b$. To derive a mathematical model, form a basic coordinate system with the z-axis positive upward. Recall the constitutive relations for each of the modeling elements, assumed here to be linear, $p_z = mV_z$, $F_k = kz_k$, and $F_b = bV_b$. In each of these elements, the associated velocity, V, or displacement, z, must be identified. The mass has a velocity, $V_z = \dot{z}$, relative to the inertial reference frame. The spring and damper have the same *relative* velocity since one end of each component is attached to the mass and the other to the base. The change in the spring length is z and the velocity is $\dot{z} - V_{base}$. However, $V_{base} = 0$ since the base is fixed, so putting this all together with Newton's second law, $m\ddot{z} = F(t) + W - kz - b\dot{z}$. A second order ordinary differential equation (ODE) is derived for this single degree of freedom (DOF) system as

$$m\ddot{z} + b\dot{z} + kz = F(t) + W$$

In this particular example, if W is left off, z is the "oscillation" about a position established by static equilibrium, $z_{static} = W/k$.

If a transfer function is desired, a simple Laplace transform leads to (assuming zero initial conditions for motion about z_{static})

$$\frac{Z(s)}{F(s)} = \frac{1}{ms^2 + bs + k}$$

The simple mass-spring-damper example illustrates that models can be readily derived for mechanical systems with direct application of kinematics and Newton's laws. As systems become more complex either due to number of bodies and geometry, or due to interaction between many types of systems (hydraulic, electromechanical, etc.), it is helpful to employ tools that have been developed to facilitate model development. In a subsequent section, multibody problems and methods of analysis are briefly discussed. It has often been argued that the utility of bond graphs can only be seen when a very complex, multi-energetic system is analyzed. This need not be true, since a system (or mechatronics) analyst can see that a consistent formulation and efficacy of causality are very helpful in analyzing many different types of physical systems. This should be kept in mind, as these basic bond graph methods are used to re-examine the simple mass-spring-damper system.

Mass-Spring-Damper: Bond Graph Approach

Figure 9.23 illustrates the development of a bond graph model for a mass-spring-damper system. In part (a), the distinct velocity points are identified and 1-junctions are used to represent them on a bond graph. Even though the base has zero velocity, and there will be no power flow into or out of that point, it is useful to identify it at this point. A relative velocity is formed using a 0-junction, and note that all bonds have sign convention applied, so at the 0-junction, $V_{mass} - V_{relative} - V_{base} = 0$, which gives, $V_{relative} = V_{mass} - V_{base}$ as required.

The model elements needed to represent the system are connected to the 1-junctions, as shown in Fig. 9.23(b). Two sources are required, one to represent the applied force (effort, S_e) due to weight, and a second to represent the fixed based velocity (a flow source, S_f). The flow source is directly attached to

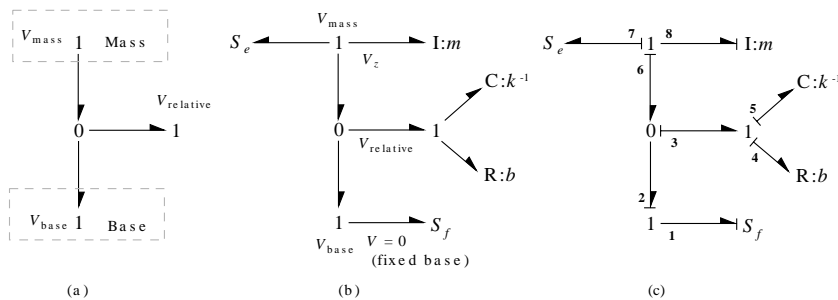



FIGURE 9.23 Basic mass-spring-damper system: (a) identifying velocity 1-junctions, (b) attaching model elements, (c) assignment of causality.

$\mathbf{C}_5$ rate law:	$\dot{z}_5$	$=$	V_5
			$V_5 = V_3$
			$V_3 = V_6 - V_2$
			$V_6 = V_8 = p_8/m_8*$
			$V_2 = V_1 = 0*$
$\mathbf{I}_8$ rate law:	$\dot{p}_8$	$=$	$-F_7 - F_6$
			$F_7 = W*$; $F_6 = F_3$
			$F_3 = F_4 + F_5$
			$F_4 = bV_4 = bV_3**$
			$F_5 = k_5 z_5*$

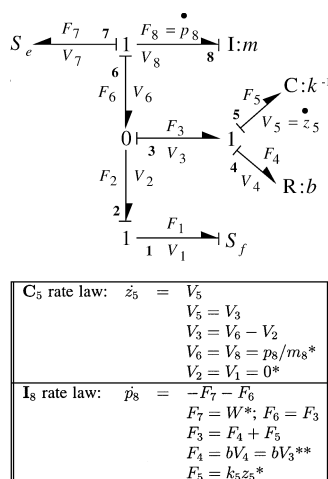



FIGURE 9.24 Equation derivation for mass-spring-damper. The '*' indicates these relations are reduced to functions of state or input. A '**' shows an intermediate variable has been reached that has elsewhere been reduced to '*'.

the 1-junction (the extra bond could be eliminated). An **I** element represents mass, a **C** represents the spring, and an **R** represents the losses in the damper. Note how the mass and the source of effort are attached to the 1-junction representing the mass velocity (the weight is always applied at that velocity). The spring and damper are attached via a power bond to the relative velocity between the mass and base.

Finally, in Fig. 9.23(c) the eight bonds are labeled and causality is assigned. First, the fixed base source fixes the causality on bond 1, specifying the velocity at the 1-junction, and thus constraining the causality of bond 2 to have effort into the 1-junction. Since bond 2 did not specify effort into the 0-junction, causality assignment should proceed to other sources, and the effort source fixes causality on bond 7. This bond does not specify the flow at the adjoining 1-junction, so at this point we could look for other specified sources. Since there are none, we assign causality to any energy-storing elements which have a preferred integral causality. The bond 8 is assigned to give the **I** element integral causality (see Table 9.7), which then specifies the velocity at the 1-junction and thus constrains bond 6. At this point, bonds 6 and 2 both specify flow into the 0-junction, so the remaining bond 3 must specify the effort. This works out well because now bond 3 specifies flow into the remaining 1-junction (the relative velocity), which specifies velocity into the **C** and **R** elements. For the **C** element, this gives integral causality.

In summary, the causality is assigned and there are no causal conflicts (e.g., two bonds trying to specify velocity into a 1-junction). Both energy-storing elements have integral causality. This indicates that the states for the **I** (mass) and **C** (spring) will contribute to the state variables of the system. This procedure assures a minimum-size state vector, which in this case is of order 2 (a 2nd-order system). Figure 9.24 shows a fully annotated bond graph, with force-velocity variables labeling each bond. The state for an **I** element is a momentum, in this case the translational momentum of the mass, p_8. For a **C** element, a

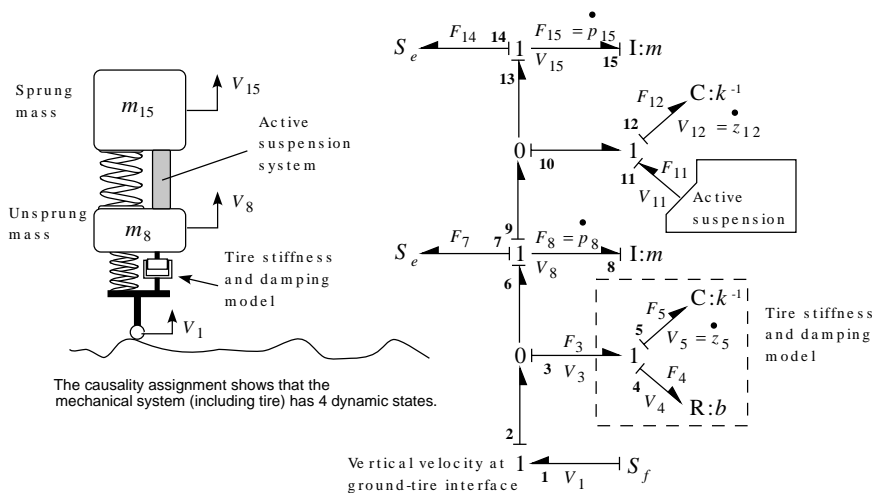



FIGURE 9.25 Example of model for vertical vibration in a quarter-car suspension model with an active suspension element. This example builds on the simple mass-spring-damper model, and shows how to integrate an actuator into a bond graph model structure.

displacement variable is the state z_5, which here represents the change in length of the spring. The state vector is $\mathbf{x}^T = [p_8, z_5]$.

A mathematical model can be derived by referring to this bond graph, focusing on the independent energy storage elements. The **rate law** (see Tables 9.4 and 9.5) for each energy storage element in integral causality constitutes one first-order ordinary differential **state equation** for this system. In order to formulate these equations, the right-hand side of each rate law must be a function only of states or inputs to the system. The process is summarized in the table of Fig. 9.24. Note that the example assumes linear constitutive relations for the elements, but it is clear in this process that this is not necessary. Of course, in some cases nonlinearity complicates the analysis as well as the modeling process in other ways.

Quarter-car Active Suspension: Bond Graph Approach

The simple mass-spring-damper system forms a basis for building more complex models. A model for the vertical vibration of a quarter-car suspension is shown in Fig. 9.25. The bond graph model illustrates the use of the mass-spring-damper model, although there are some changes required. In this case, the base is now moving with a velocity equal to the vertical velocity of the ground-tire interface (this requires knowledge of the terrain height over distance traveled as well as the longitudinal velocity of the vehicle). The power direction has changed on many of the bonds, with many now showing positive power flowing from the ground up into the suspension system.

The active suspension system is isolated to further illustrate how bond graph modeling promotes a modular approach to the study of complex systems. Most relevant is that the model identifies the required causal relation at the interface with the active suspension, specifying that the relative velocity is a causal input, and force is a causal output of the active suspension system. The active force is exerted in an equal and opposite fashion onto the sprung and unsprung mass elements.

The causality assignment identifies four states (two momentum states and two spring displacement states). Four first-order state equations can be derived using the rate laws of each of the independent energy-storing elements ($\mathbf{C}_5$, $\mathbf{I}_8$, $\mathbf{C}_{12}$, $\mathbf{I}_{15}$). At this point, depending on the goals of the analysis, either the nonlinear equations could be derived (which might include an active suspension force that depends on the velocity input), or a linearized model could be developed and impedance methods applied to derive a transfer function directly.

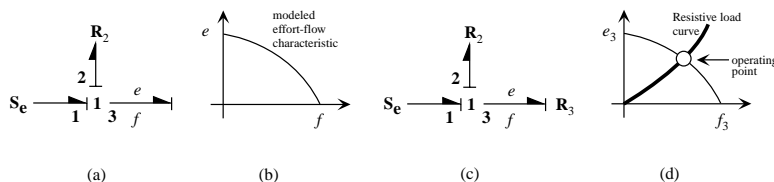



FIGURE 9.26 Algebraic loop in a simple source-load model.

Note on Some Difficulties in Deriving Equations

There are two common situations that can lead to difficulties in the mathematical model development. These issues will arise with any method, and is not specific to bond graphs. Both lead to a situation that may require additional algebraic manipulation in the equation derivation, and it may not be possible to accomplish this in closed form. There are also some ways to change the model in order to eliminate these problems, but this could introduce additional problems. The two issues are (1) derivative causality, and (2) algebraic loops. Both of these can be detected during causality assignment, so that a problem can be detected before too much time has been spent.

The occurence of derivative causality can be described in bond graph terms using Table 9.7. The issue is one in which the state of an energy-storing element (**I** or **C**) is dependent on the system to which it is attached. This might not seem like a problem, particularly since this implies that no differential equation need be solved to find the state. It is necessary to see that there is still a need to compute the back-effect that the system will feel in forcing the element into a given state. For example, if a mass is to be driven by a velocity, V, then it is clear that we know the energy state, $p = mV$, so all is known. However, there is an inertial force computed as $\dot{p} = m\dot{V} = ma$. Many times, it is possible to resolve this problem by performing the algebraic manipulations required to include the effect of this element (difficulty depends on complexity of the system). Sometimes, these dependent states arise because the system is not modeled in sufficient detail, and by inserting a compliance between two gears, for example, the dependence is removed. This might solve the problem, costing only the introduction of an additional state. A more serious drawback to this approach would occur if the compliance was actually very small, so that numerical stiffness problems are introduced (with modern numerical solver routines, even this problem can be tolerated). Yet another way to resolve the problem of derivative causality in mechanical systems is to employ a Lagrangian approach for mechanical system modeling. This will be discussed in section 9.7.

Another difficulty that can arise in developing solvable systems of equations is the presence of an algebraic loop. Algebraic loops are relatively easy to generate, especially in a block diagram modeling environment. Indeed, it is often the case that algebraic loops arise because of modeling decisions, and in this way a bond graph's causality provides quick feedback regarding the system solvability. Algebraic loops imply that there is an arbitrary way to make computations in the model, and in this way they reveal themselves when an arbitrary decision must be made in assigning causality to an **R** element.[3]

As an example, consider the basic model of a Thevenin source in Fig. 9.26(a). This model uses an effort source and a resistive element to model an effort-flow (steady-state) characteristic curve, such as a motor or engine torque-speed curve or a force-velocity curve for a linear actuator. A typical characteristic is shown in Fig. 9.26(b). When a resistive load is attached to this source as shown in Fig. 9.26(c), the model is purely algebraic. When the causality is assigned, note that after applying the effort causality on bond 1, there are two resistive elements remaining. The assignment of causality is arbitrary. The solution

[3]The arbitrary assignment on an **R** element is not unlike the arbitrariness in assigning integral or derivative causality to energy-storing elements. An "arbitrary" decision to assign integral causality on an energy-storing element leads to a requirement that we solve a *differential* equation to find a state of interest. In the algebraic loop, a similar arbitary decision to assign a given causality on an **R** element implies that at least one *algebraic* equation must be solved along with any other system equations. In other words, the system is described by differential algebraic equations (DAEs).

requires analytically solving algebraic relations for the operating point, or by using a graphical approach as shown in Fig. 9.26(d).

This is a simple example indicating how algebraic loops are detected with a bond graph, and how the solution requires solving algebraic relations. In complex systems, this might be difficult to achieve. Sometimes it is possible to introduce or eliminate elements that are "parasitic," meaning they normally would be neglected due to their relatively small effect. However, such elements can relieve the causal bind. While this might resolve the problem, as in the case of derivative causality there are cases where such a course could introduce numerical stiffness problems. Sometimes a solution is reached by using energy methods to resolve some of these problems, as shown in the next section.

9.5 Energy Methods for Mechanical System Model Formulation

This section describes methods for using energy functions to describe basic energy-storing elements in mechanical systems, as well as a way to describe collections of energy-storing elements in multiport fields. Energy methods can be used to simplify model development, providing the means for deriving constitutive relations, and also as a basis for eliminating dependent energy storage (see last section). The introduction of these methods provides a basis for introducing the Lagrange equations in section 9.7 as a primary approach for system equation derivation or in combination with the bond graph formulation.

Multiport Models

The energy-storing and resistive models introduced in section 9.3 were summarized in Tables 9.2, 9.4, and 9.5 as multiport elements. In this section, we review how multiport elements can be used in modeling mechanical systems, and outline methods for deriving the constitutive relations. Naturally, these methods apply to the single-port elements as well.

An example of a **C** element with two-ports was shown in Fig. 9.12 as a model for a cantilevered beam that can have both translational and rotational deflections at its tip. A 2-port is required in this model because there are two independent ways to store potential energy in the beam. A distinguishing feature in this example is that the model is based on relationships between efforts and displacement variables (for this case of a capacitive element). Multiport model elements developed in this way are categorized as explicit fields to distinguish them from implicit fields [17]. Implicit fields are formed by assembling energy-storing 1-port elements with junction structure (i.e., 1, 0, and TF elements) to form multiport models.

Explicit fields are often derived using physical laws directly, relying on an understanding of how the geometric and material properties affect the basic constitutive relation between physical variables. Geometry and material properties always govern the parametric basis of all constitutive relations, and for some cases these properties may themselves be functions of state. Indeed, these cases require the multiport description, which finds extensive use in modeling of many practical devices, especially sensors and actuators. Multiport models should follow a strict energetic basis, as described in the following.

Restrictions on Constitutive Relations

Energy-storing multiports must follow two basic restrictions, which are also useful in guiding the derivation of energetically-correct constitutive relations. The definition of the energy-storing descriptions summarized in Tables 9.4 and 9.5 specifies that there exists an energy state function, $E = E(\mathbf{x})$, where $\mathbf{x}$ is either a generalized displacement, $\mathbf{q}$, for capacitive (**C**) elements or a generalized momentum, $\mathbf{p}$, for inertive (**I**) elements. For the multiport energy-storing element, the specification requires the following specifications [2,3].

1. There exists a rate law, $\dot{x}_i = u_i$, where u_i as input specifies integral causality on port i.
2. The energy stored in a multiport is determined by

$$E(\mathbf{x}) = \int \sum_{i=1}^{n} y_i \, dx_i \tag{9.5}$$

3. A first restriction on a multiport constitutive relation requires that the causal output at any port is given by

$$y_i = \Phi_{si}(\mathbf{x}) = \frac{\partial E(\mathbf{x})}{\partial x_i} \tag{9.6}$$

where $F_{si}()$ is a single-valued function.

4. A second restriction on a multiport constitutive relation requires that the constitutive relations obey Maxwell reciprocity, or

$$\frac{\partial y_i}{\partial x_j} = \frac{\partial^2 E(\mathbf{x})}{\partial x_j \partial x_i} = \frac{\partial y_j}{\partial x_i} \tag{9.7}$$

Deriving Constitutive Relations

The first restriction on the constitutive relations, Eq. (9.6), establishes how constitutive relations can be derived for a multiport if an energy function can be formulated. This restriction forms the basis for a method used in many practical applications to find constitutive relationships from energy functions (e.g., strain-energy, electromechanics, etc.). In these methods, it is assumed that at least one of the constitutive relations for an energy-storing multiport is given. Then, the energy function is formed using Eq. (9.5) where, after interchanging the integral and sum,

$$(\mathbf{x}) = \sum_{i=1}^{n} \int y_i\, dx_i = \int y_1\, dx_1 + {}^{\circ} + \int y_n d \tag{9.8}$$

Presume that y_1 is a known function of the states, $y_1 = \Phi_{si}(\mathbf{x})$. Since the element is conservative, any energetic state can be reached via a convenient path where $dx_i = 0$ for all i except $i = 1$. This allows the determination of $E(\mathbf{x})$.

To illustrate, consider the simple case of a rack and pinion system, shown in Fig. 9.27. The pinion has rotational inertia, J, about its axis of rotation, and the rack has mass, m. The kinetic co-energy is easily formulated here, considering that the pinion angular velocity, ω, and the rack velocity, V, are constrained by the relationship $V = R\omega$, where R is the pinion base radius. If this basic subsystem is modeled directly, it will be found that one of the inertia elements (pinion, rack) will be in derivative causality. Say, it is desired to connect to this system through the rotational port, T - ω. To form a single-port **I** element that includes the rack, form the kinetic co-energy as $T = T(\omega, V) = J\omega^2/2 + mV^2/2$. Use the constraint relation to write, $T = T(\omega) = (J + mR^2)\omega^2/2$. To find the constitutive relation for this 1-port rotational **I** element, let $h = \partial T(\omega)/\partial\omega = (J + mR^2)\omega$, where we can now define an equivalent rotational inertia as $J_{eq} = J + mR^2$.

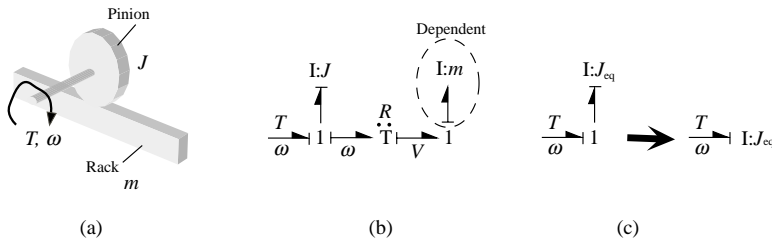



FIGURE 9.27 (a) Rack and pinion subsystem with torque input. (b) Direct model, showing dependent mass. (c) Equivalent model, derived using energy principles.

The rack and pinion example illustrates a basic method for relieving derivative causality, which can be used to build basic energy-storing element models. Some problems might arise when the kinetic co-energy depends on system configuration. In such a case, a more systematic method employing Lagrange's equations may be more suitable (see Section 9.7).

The approach described here for deriving constitutive relations is similar to Castigliano's theorom [6,9]. Castigliano's theorem relies on formulation of a strain-energy function in terms of the forces or moments, and as such employs a potential co-energy function. Specifically, the results lead to displacements (translational, rotational) as functions of efforts (forces, torques). As in the case above, these functions are found by taking partial derivatives of the co-energy with respect to force or moment. Castigliano's theorem is especially well-suited for finding force-displacment functions for curved and angled beam structures (see [6]).

Formulations using energy functions to derive constitutive relations are found in other application areas, and some references include Lyshevski [21] for electromechanics, and Karnopp, Margolis, and Rosenberg [17] for examples and applications in the context of bond graph modeling.

Checking the Constitutive Relations

The second restriction on the constitutive relations, Eq. (9.7), provides a basis for testing or checking if the relationships are correct. This is a reciprocity condition that provides a check for energy conservation in the energy-storing element model, and a quick check for linear mechanical systems shows that either the inertia or stiffness matrix must be symmetrical.

Recall the example of the 2-port cantilevered beam, shown again in Fig. 9.12. For small deflections, the total tip translational and angular deflections due to a tip force and torque can be added (using flexibility influence coefficients), which can be expressed in matrix form,

$$\begin{bmatrix} x \\ \theta \end{bmatrix} = \frac{1}{EI}\begin{bmatrix} \frac{1}{3}l^3 & \frac{1}{2}l^2 \\ \frac{1}{2}l^2 & l \end{bmatrix}\begin{bmatrix} F \\ T \end{bmatrix} = \mathbf{C}\begin{bmatrix} F \\ T \end{bmatrix} = \mathbf{K}^{-1}\begin{bmatrix} F \\ T \end{bmatrix}$$

where **C** and **K** are the compliance and stiffness matrices, respectively. This constitutive relation satisfies the Maxwell reciprocity since, $\partial x/\partial T = \partial\theta/\partial F$. This 2-port **C** element is used to model the system shown in Fig. 9.28(a), which consists of a bar-bell rigidly attached to the tip of the beam. Under small deflection, a bond graph shown in Fig. 9.28(b) is assembled. Causality applied to this system reveals that each port of the 2-port **C** element has integral causality. On a multiport energy storing element, each port is independently assigned causality following the same rules as for 1-ports. It is possible that a multiport could have a mixed causality, where some of the ports are in derivative causality. If a multiport has mixed causality, part of the state equations will have to be inverted. This algebraic difficulty is best avoided by trying to assign integral causality to all multiport elements in a system model if possible.

In the present example, causality assignment on the **I** elements is also integral. In all, there are four independent energy-storing elements, so there are four state variables, $\mathbf{x} = [x, \theta, p, h]^\top$. Four state equations can be derived using the rate laws indicated in Fig. 9.28.

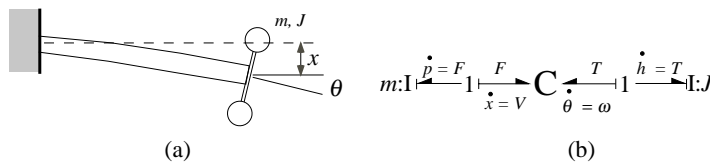



FIGURE 9.28 Model of beam rigidly supporting a bar- or dumb-bell: (a) schematic, (b) bond graph model using a 2-port **C** to represent beam. Dumb-bell is represented by translational mass, *m*, and rotational inertia, *J*.

9.6 Rigid Body Multidimensional Dynamics

The modeling of bodies in mechanical systems presumes adoption of a “rigid body” that can involve rotation as well as translation, and in this case the dynamic properties are more complex than those for a point mass. In earlier sections of this chapter, a simple rigid body has already been introduced, and it is especially useful for a large class of problems with rotation about a single fixed axis.

In the rigid body, the distance between any two elements of mass within a body is a constant. In some cases, it is convenient to consider a continuous distribution of mass while in others a system of discrete mass particles rigidly fixed together helps conceptualize the problem. In the latter, the rigid body properties can be found by summing over all the discrete particles, while in the continuous mass concept an integral formulation is used. Either way, basic concepts can be formulated and relations derived for use in rigid body dynamic analysis. Finally, the modeling in most engineering systems is restricted to classical Newtonian mechanics, where the linear velocity–momentum relation holds (so energy and coenergy are equal).

Kinematics of a Rigid Body

In this section, a brief overview is given of three-dimensional motion calculations for a rigid body. The focus here is to present methods for analyzing rotation of a rigid body about a fixed axis and methods for analyzing relative motion of a rigid body using translating and rotating axes. These concepts introduce the basis for understanding more complex formulations. While vector descriptions (denoted using an arrow over the symbol, $\vec{a}$) are useful for understanding basic problems, more complex multibody systems usually adopt a matrix formulation. The presentation here is brief and included for reference. A more extensive discussion and examples can be found in introductory dynamics textbooks (e.g., [23]), where a separate discussion is usually given on the special case of plane motion.

Rotation of a Body About a Fixed Point

Basic concepts are introduced here in relation to rotation of a rigid body about a fixed point. This basic motion specifies that any point on the body lies on the surface of a sphere with a radius centered at the fixed point. The body can be said to have spherical motion.

Euler’s Theorem. Euler’s theorem states that any displacement of a body in spherical motion can be expressed as a rotation about a line that passes through the center of the spherical motion. This axis can be referred to as the orientational axis of rotation [26]. For example, two rotations about different axes passing through a fixed point of rotation are equivalent to a single resultant rotation about an axis passing through that point.

Finite Rotations. If the rotations used in Euler’s theorem are finite, the order of application is important because finite rotations do not obey the law of vector addition.

Infinitesimal Rotations. Infinitesimally small rotations can be added vectorially in any manner, and these are generally considered when defining rigid body motions.

Angular Velocity. A body subjected to rotation $d\vec{\theta}$ about a fixed point will have an angular velocity $\vec{w}$ defined by the time derivative $d\vec{\theta}/dt$, in a direction collinear with $d\vec{\theta}$. If the body is subjected to two component angular motions that define w_1 and w_2, then the body has a resultant angular velocity, $w = \vec{w}_1 + \vec{w}_2$.

Angular Acceleration. A body’s angular acceleration is found from the time derivative of the angular velocity, $\vec{\alpha} = \dot{\vec{\omega}}$, and in general the acceleration is not collinear with velocity.

Motion of Points in the Body. Given ω, the velocity of a point on the body is $\vec{v} = \vec{\omega} \times \vec{r}$, where $\vec{r}$ is a position vector to the point as measured relative to the fixed point of rotation. The acceleration of a point on the body is then, $\vec{\alpha} = \vec{\alpha} \times \vec{r} + \vec{\omega} \times (\vec{\omega} \times \vec{r})$.

Relating Vector Time Derivatives in Coordinate Systems

It is often the case that we need to determine the time rate of change of a vector such as $\vec{A}$ in Fig. 9.29 relative to different coordinate systems. Specifically, it may be easier to determine $\vec{A}$ in x_a, y_a, z_a, but we

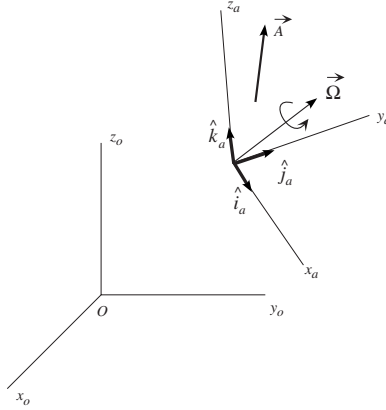



FIGURE 9.29 Often it is necessary to find the time derivative of vector $\vec{A}$ relative to a axes, x_o, y_o, z_o, given its value in the translating-rotating system x_a, y_a, z_a.

need to find its value in x_o, y_o, z_o. The vector $\vec{A}$ is expressed in the axes x_a, y_a, z_a using the unit vectors shown as

$$\vec{A} = A_x\hat{i}_a + A_y\hat{j}_a + A_z\hat{k}_a$$

To find the time rate of change, we identify that in the moving reference the time derivative of $\vec{A}$ is

$$\left(\frac{d\vec{A}}{dt}\right)_a = \frac{dA_x}{dt}\hat{i}_a + \frac{dA_y}{dt}\hat{j}_a + \frac{dA_z}{dt}\hat{k}_a$$

Relative to the x_o, y_o, z_o axes, the direction of the unit vectors $\hat{i}_a$, $\hat{j}_a$, and $\hat{k}_a$ change only due to rotation Ω, so,

$$\frac{d\vec{A}}{dt} = \left(\frac{d\vec{A}}{dt}\right) + A_x\frac{d\hat{i}_a}{dt} + A_y\frac{d\hat{j}_a}{dt} + A_z\frac{d\hat{k}_a}{dt}$$

$$\frac{d\hat{i}_a}{dt} = \vec{\Omega}\times\hat{i}_a, \quad \frac{d\hat{j}_a}{dt} = \vec{\Omega}\times\hat{j}_a, \quad \frac{d\hat{k}_a}{dt} = \vec{\Omega}\times\hat{k}_a$$

then,

$$\frac{d\vec{A}}{dt} = \left(\frac{d\vec{A}}{dt}\right)_a + \vec{\Omega}\times\vec{A} \tag{9.9}$$

This relationship is very useful not only for calculating derivatives, as derived here, but also for formulating basic bond graph models. This is shown in the section titled "Rigid Body Dynamics."

Motion of a Body Relative to a Coordinate System

Translating Coordinate Axes

The origin of a set of axes x_a, y_a, z_a is fixed in a rigid body at A as shown in Fig. 9.30(a), and translates without rotation relative to the axes x_o, y_o, z_o with known velocity and acceleration. The rigid body is subjected to angular velocity $\vec{\omega}$ and angular acceleration $\vec{\alpha}$ in three dimensions.

Motion of Point B Relative to A. The motion of point B relative to A is the same as motion about a fixed point, so $\vec{v}_{B/A} = \vec{\omega}\times\vec{v}_{B/A}$, and $\vec{a}_{B/A} = \vec{\alpha}\times\vec{r}_{B/A} + \vec{\omega}\times(\vec{\omega}\times\vec{r}_{B/A})$.

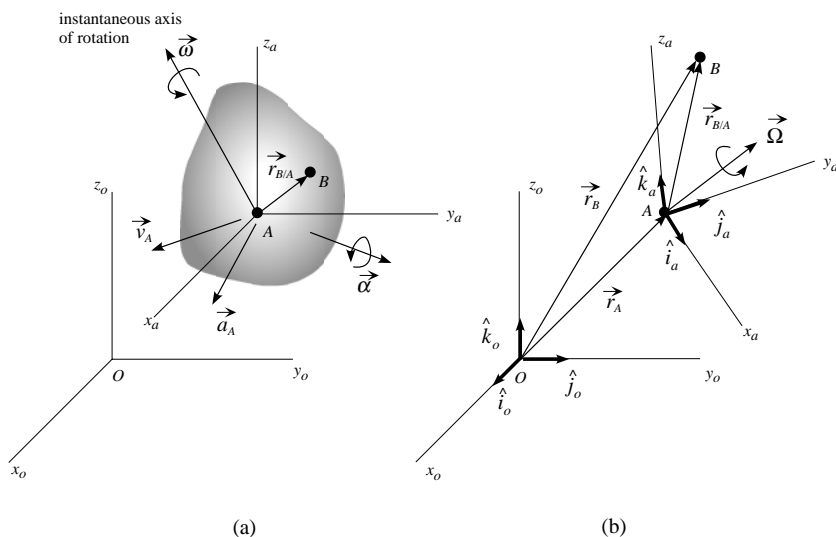



FIGURE 9.30 General rigid body motion: (a) rigid body with translating coordinate system, (b) translating and rotating coordinate system.

Motion of Point *B* Relative to *O*. For translating axes with no rotation, the velocity and acceleration of point B relative to system 0 is simply, $\vec{v}_B = \vec{v}_A + \vec{v}_{B/A}$ and $\vec{a}_B = \vec{a}_A + \vec{a}_{B/A}$ respectively, or,

$$\vec{v}_B = \vec{v}_A + \vec{\omega} \times \vec{r}_{B/A} \tag{9.10}$$

$$\vec{a}_B = \vec{a}_A + \vec{\alpha} \times \vec{r}_{B/A} + \vec{\omega} \times (\vec{\omega} \times \vec{r}_{B/A}) \tag{9.11}$$

Translating and Rotating Coordinate Axes

A general way of describing the three-dimensional motion of a rigid body uses a set of axes that can translate and rotate relative to a second set of axes, as illustrated in Fig. 9.30(b). Position vectors specify the locations of points A and B on the body relative to x_o, y_o, z_o, and the axes x_a, y_a, z_a have angular velocity $\vec{\Omega}$ and angular acceleration $\vec{\Omega}$. With the position of point B given by

$$\vec{r}_B = \vec{r}_A + \vec{r}_{B/A} \tag{9.12}$$

the velocity and acceleration are found by direct differentiation as

$$\vec{v}_B = \vec{v}_A + \vec{\Omega} \times \vec{r}_{B/A} + (v_{B/A})_a \tag{9.13}$$

and

$$_B = \vec{a}_A + \dot{\vec{\Omega}} \times \vec{r}_{B/A} + \Omega \text{ ¥ } (\Omega \times \vec{r}_{B/A}) + 2\Omega \times (v_{B/A})_a + (\vec{a}_{B/A}) \tag{9.14}$$

where $(v_{B/A})_a$ and $(a_{B/A})_a$ are the velocity and acceleration, respectively, of B relative to A in the x_a, y_a, z_a coordinate frame.

These equations are applicable to plane motion of the rigid body for which the analysis is simplified since $\vec{\Omega}$ and $\vec{\Omega}$ have a constant direction. Note that for the three-dimensional case, $\vec{\Omega}$ must be computed by using Eq. (9.9).

Matrix Formulation and Coordinate Transformations

A vector in three-dimensional space characterized by the right-handed reference frame x_a, y_a, z_a, $\vec{A} = A_x\hat{i}_a + A_y\hat{j}_a + A_z\hat{k}_a$, can be represented as an ordered triplet,

$$\vec{A} = \begin{bmatrix} A_x \\ A_y \\ A_z \end{bmatrix}_a = \begin{bmatrix} A_x & A_y & A_z \end{bmatrix}_a^{\mathsf{T}}$$

where the elements of the column vector represent the vector projections on the unit axes. Let $\underline{A}_a$ denote the column vector relative to the axes x_a, y_a, z_a. It can be shown that the vector $\vec{A}$ can be expressed in another right-handed reference frame x_b, y_b, z_b, by the transformation relation

$$\underline{A}_b = \bar{\underline{C}}_{ab}\underline{A}_a \tag{9.15}$$

where $\bar{\underline{C}}_{ab}$ is a 3×3 matrix,

$$\bar{\underline{C}}_{ab} = \begin{bmatrix} cx_a x_b & cx_a y_b & cx_a z_b \\ cy_a x_b & cy_a y_b & cy_a z_b \\ cz_a x_b & cz_a y_b & cz_a z_b \end{bmatrix} \tag{9.16}$$

The elements of this matrix are the cosines of the angles between the respective axes. For example, $cz_a y_b$ is the cosine of the angle between z_a and y_b. This is the rotational transformation matrix and it must be orthogonal, or

$$C_{ab}^{-1} = C_{ab}^{\mathsf{T}} = C_{ba}$$

and for right-handed systems, let $C_{ab} = +1$.

Angle Representations of Rotation

The six degrees of freedom needed to describe general motion of a rigid body are characterized by three degrees of freedom each for translation and for rotation. The focus here is on methods for describing rotation.

Euler's theorem (11) confirms that only three parameters are needed to characterize rotation. Two parameters define an axis of rotation and another defines an angle about that axis. These parameters define three positional degrees of freedom for a rigid body. The three rotational parameters help construct a rotation matrix, $\bar{\underline{C}}$. The following discussion describes how the rotation matrix, or direction cosine matrix, can be formulated.

General Rotation. Unit vectors for a system a, $\hat{u}_a$, are said to be carried into b, as $\hat{u}_b = \bar{\underline{C}}_{ba}\hat{u}_a$. It can be shown that a direction cosine matrix can be formulated by [30]

$$\bar{\underline{C}} = \underline{\lambda}\underline{\lambda}^{\mathsf{T}} + (\bar{\underline{E}} - \underline{\lambda}\underline{\lambda}^{\mathsf{T}})\cos\psi - \bar{\underline{S}}(\lambda)\sin\psi \tag{19.17}$$

where $\bar{\underline{E}}$ is the identity matrix, and $\underline{\lambda}$ represents a unit vector, $\underline{\lambda} = [\lambda_1, \lambda_2, \lambda_3]^{\mathsf{T}}$, which is parallel to the axis of rotation, and ψ is the angle of rotation about that axis [30]. In this relation, $\bar{\underline{S}}(\underline{\lambda})$ is a **skew-symmetric matrix**, which is defined by the form

$$\bar{\underline{S}}(\underline{\lambda}) = \begin{bmatrix} 0 & -\lambda_3 & \lambda_2 \\ \lambda_3 & 0 & -\lambda_1 \\ -\lambda_2 & \lambda_1 & 0 \end{bmatrix}$$

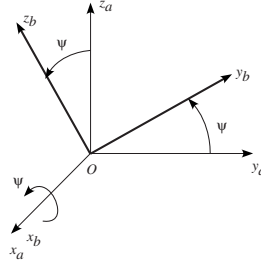



FIGURE 9.31 An elementary rotation by angle ϕ about axis x.

The matrix elements of $\underline{\bar{C}}$ can be found by expanding the relation given above, using $\underline{\bar{S}}(\lambda)$, to give

$$\underline{\bar{C}} = \begin{bmatrix} (1-\cos\psi)\lambda_1^2 + \cos\psi & (1-\cos\psi)\lambda_1\lambda_2 + \lambda_3\sin\psi & (1-\cos\psi)\lambda_1\lambda_3 + \lambda_2\sin\psi \\ (1-\cos\psi)\lambda_2\lambda_1 + \lambda_3\sin\psi & (1-\cos\psi)\lambda_2^2 + \cos\psi & (1-\cos\psi)\lambda_2\lambda_3 + \lambda_1\sin\psi \\ (1-\cos\psi)\lambda_3\lambda_1 + \lambda_2\sin\psi & (1-\cos\psi)\lambda_3\lambda_2 + \lambda_1\sin\psi & (1-\cos\psi)\lambda_3^2 + \cos\psi \end{bmatrix} \quad (9.18)$$

The value of this formulation is in identifying that there are formally defined principle axes, characterized by the $\underline{\lambda}$, and angles of rotation, ψ, that taken together define the body orientation. These rotations describe classical angular variables formed by elementary (or principle) rotations, and it can be shown that there are two cases of particular and practical interest, formed by two different axis rotation sequences.

Elementary Rotations. Three elementary rotations are formed when the rotation axis (defined by the eigenvector) coincides with one of the base vectors of a defined coordinate system. For example, letting $\underline{\lambda} = [1,\ 0,\ 0]^{\mathsf{T}}$ define an axis of rotation x, as in Fig. 9.31, with an elementary rotation of ϕ gives the rotation matrix,

$$\underline{\bar{C}}_{x,\phi} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & \sin\phi \\ 0 & -\sin\phi & \cos\phi \end{bmatrix}$$

The two elementary rotations about the other two axes, y and z, are

$$\underline{\bar{C}}_{y,\theta} = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix} \quad \text{and} \quad \underline{\bar{C}}_{z,\psi} = \begin{bmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

These three elementary rotation matrices can be used in sequence to define a direction cosine matrix, for example,

$$\underline{\bar{C}} = \underline{\bar{C}}_{z,\psi}\underline{\bar{C}}_{y,\theta}\underline{\bar{C}}_{x,\phi}$$

and the elementary rotations and the direction cosine matrix are all orthogonal; i.e.,

$$\underline{\bar{C}}\,\underline{\bar{C}}^{\mathsf{T}} = \underline{\bar{C}}^{\mathsf{T}}\underline{\bar{C}} = \underline{\bar{E}}$$

where $\underline{\bar{E}}$ is the identity matrix. Consequently, the inverse of the rotation or coordinate transformation matrix can be found by $\underline{\bar{C}}^{-1} = \underline{\bar{C}}^{\mathsf{T}}$.

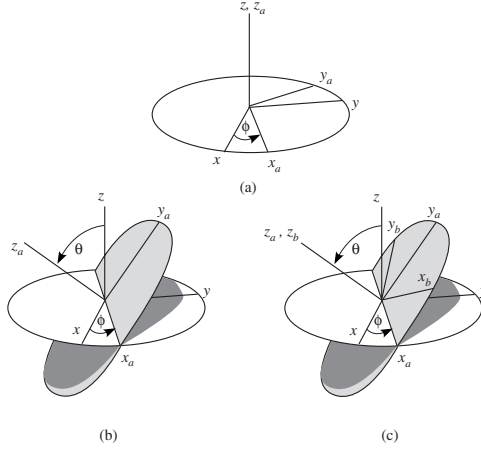



FIGURE 9.32 The rotations defining the Euler angles (adapted from Goldstein [11]).

It can be shown that there exist two sequences that have independent rotation sequences, and these lead to the well known Euler angle and Tait-Bryan or Cardan angle rotation descriptions [30].

Euler Angles. Euler angles are defined by a specific rotation sequence. Consider a right-handed axes system defined by the base vectors, x, y, z, as shown in Fig. 9.32(a). The rotation sequence of interest involves rotations about the axes in the following sequence: (1) ϕ about z, (2) θ about x_a, then (3) ψ about z_b. This set of rotation sequences is defined by the elementary rotation matrices,

$$\underline{\overline{C}}_{z,\phi} = \begin{bmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \underline{\overline{C}}_{x_a,\theta} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{bmatrix}, \quad \underline{\overline{C}}_{z_b,\psi} = \begin{bmatrix} \cos\psi & \sin\psi & 0 \\ -\sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

where the subscript on each $\underline{\overline{C}}$ denotes the axis and angle of rotation. Using these transformations relates the quantity $\underline{A}$ in x, y, z to $\underline{A}_b$ in x_b, y_b, z_b, or

$$\underline{A}_b = \underline{\overline{C}}_{\text{Euler}}\,\underline{A} = \underline{\overline{C}}_{z_b,\psi}\underline{\overline{C}}_{x_a,\theta}\underline{\overline{C}}_{z,\phi}\,\underline{A}$$

where $\underline{\overline{C}}_{\text{Euler}}$ is given by

$$\underline{\overline{C}}_{\text{Euler}} = \begin{bmatrix} \cos\psi\cos\phi - \sin\psi\cos\theta\sin\phi & \cos\psi\sin\phi + \sin\psi\cos\theta\cos\phi & \sin\psi\sin\theta \\ -\sin\psi\cos\phi - \cos\psi\cos\theta\sin\phi & -\sin\psi\sin\phi + \cos\psi\cos\theta\cos\phi & \cos\psi\sin\theta \\ \sin\theta\sin\phi & -\sin\theta\cos\phi & \cos\theta \end{bmatrix} \tag{9.19}$$

Since $\underline{\overline{C}}_{\text{Euler}}$ is orthogonal, transforming between the two coordinate systems is relatively easy since the inverse can be found simply by the transpose of Eq. (9.19).

In some applications, it is desirable to derive the angles given the direction cosine matrix. So, if the (3,3) element of $\underline{\overline{C}}_{\text{Euler}}$ is given, then θ is easily found, but there can be difficulties in discerning small angles. Also, if θ goes to zero, there is a singularity in solving for ϕ and ψ, so determining body orientation becomes difficult. The problem also makes itself known when transforming angular velocities between the coordinate systems. If the problem at hand avoids this case (i.e., θ never approaches zero), then Euler angles are a viable solution. Many applications that cannot tolerate this problem adopt other representations, such as the Euler parameters to be discussed later.

In classical rigid body dynamics, ϕ is called the *precession angle,* θ is the *nutation angle*, and ψ is the *spin angle*. The relationship between the time derivative of the Euler angles, $\dot{\underline{\varphi}} = [\dot{\phi}, \dot{\theta}, \dot{\psi}]^T$, and the body angular velocity, $\underline{\omega} = [\omega_x, \omega_y, \omega_z]_b^T$, is given by [11]

$$\underline{\omega}_b = \underline{\overline{T}}(\underline{\varphi})\dot{\underline{\varphi}} \tag{9.20}$$

where the transformation matrix, $\underline{\overline{T}}(\underline{\varphi})$, is given by

$$\underline{\overline{T}}(\underline{\varphi}) = \begin{bmatrix} \sin\theta\sin\psi & \cos\psi & 0 \\ \sin\theta\cos\psi & -\sin\psi & 0 \\ \cos\theta & 0 & 1 \end{bmatrix}$$

Note here again that $\underline{\overline{T}}(\varphi)$ will become singular at $\theta = \pm\pi/2$.

Tait-Bryan or Cardan Angles. The Tait-Bryan or Cardan angles are formed when the three rotation sequences each occur about a different axis. This is the sequence preferred in flight and vehicle dynamics. Specifically, these angles are formed by the sequence: (1) ϕ about z (yaw), (2) θ about y_a (pitch), and (3) ϕ about the final x_b axis (roll), where a and b denote the second and third stage in a three-stage sequence or axes (as used in the Euler angle description). These rotations define a transformation,

$$\underline{A}_b = \underline{\overline{C}}\,\underline{A} = \underline{\overline{C}}_{x_b,\psi}\underline{\overline{C}}_{y_a,\theta}\underline{\overline{C}}_{z,\phi}\,\underline{A}$$

where

$$\underline{\overline{C}}_{z,\phi} = \begin{bmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \underline{\overline{C}}_{y_a,\theta} = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix}, \quad \underline{\overline{C}}_{x_b,\theta} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\psi & \sin\psi \\ 0 & -\sin\psi & \cos\psi \end{bmatrix}$$

and the final coordinate transformation matrix for Tait-Bryan angles is

$$\underline{\overline{C}}_{\text{Tait-Bryan}} = \begin{bmatrix} \cos\theta\cos\phi & \cos\theta\sin\phi & -\sin\theta \\ \sin\psi\sin\theta\cos\phi - \cos\psi\sin\phi & \sin\psi\sin\theta\sin\phi + \cos\psi\cos\phi & \cos\theta\sin\psi \\ \cos\psi\sin\theta\cos\phi + \sin\psi\sin\phi & \cos\psi\sin\theta\sin\phi - \sin\psi\cos\phi & \cos\theta\cos\psi \end{bmatrix} \tag{9.21}$$

A linearized form of $\underline{\overline{C}}_{\text{Trait-Bryan}}$ gives a form preferred to that derived for Euler angles, making it useful in some forms of analysis and control. There remains the problem of a singularity, in this case when θ approaches $\pm\pi/2$.

For the Tait-Bryan angles, the transformation matrix relating $\dot{\underline{\varphi}}$ to $\underline{\omega}_b$ is given by

$$\underline{\overline{T}}(\underline{\varphi}) = \begin{bmatrix} -\sin\theta & 0 & 1 \\ \cos\theta\sin\psi & \cos\psi & 0 \\ \cos\theta\cos\psi & -\sin\psi & 0 \end{bmatrix}$$

which becomes singular at $\theta = 0, \pi$.

Euler Parameters and Quaternions

The degenerate conditions in coordinate transformations for Euler and Tait-Bryan angles can be avoided by using more than a minimal set of parameterizing variables (beyond the three angles). The most notable

set are referred to as Euler parameters, which are unit quaternions. There are many other possibilities, but this four-parameter method is used in many areas, including spacecraft/flight dynamics, robotics, and computational kinematics and dynamics. The term "quaternion" was coined by Hamilton in about 1840, but Euler himself had devised the use of Euler parameters 70 years before. Quaternions are discussed by Goldstein [11], and their use in rigid body dynamics and attitude control dates back to the late 1950s and early 1960s [13,24]. Application of quaternions is common in control applications in aerospace applications [38] as well as in ocean vehicles [10]. More recently (past 20 years or so), these methods have found their way into motion and control descriptions for robotics [34] and computational kinematics and dynamics [14,25,26]. An overview of quaternions and Euler parameters is given by Wehage [37]. Quaternions and rotational sequences and their role in a wide variety of applications areas, including sensing and graphics, are the subject of the book by Kuipers [19]. These are representative references that may guide the reader to an application area of interest where related studies can be found. In the following only a brief overview is given.

Quaternion. A quaternion is defined as the sum of a scalar, q_0, and a vector, $\vec{q}$, or,

$$q = q_0 + \vec{q} = q_0 + q_1\hat{i} + q_2\hat{j} + q_3\hat{k}$$

A specific algebra and calculus exists to handle these types of mathematical objects [7,19,37]. The conjugate is defined as $q = q_0 - \vec{q}$.

Euler Parameters. Euler parameters are normalized (unit) quaternions, and thus share the same properties, algebra and calculus. A principal eigenvector of rotation has an eigenvalue of 1 and defines the Euler axis of rotation (see Euler's theorem discussion and [11]), with angle of rotation α. Let this eigenvector be $\underset{\sim}{e} = [e_1,\ e_2,\ e_3]^{\mathrm{T}}$. Recall from Eq. (9.17), the direction cosine matrix is now

$$\underset{=}{C} = \underline{e}\underline{e}^{\mathrm{T}} + (I - \underline{e}\underline{e}^{\mathrm{T}})\cos\alpha - \underset{=}{S}(\underset{\sim}{e})\sin\alpha$$

where $\underset{=}{S}(\underset{\sim}{e})$ is a skew-symmetric matrix. The Euler parameters are defined as

$$\underset{\sim}{q} = \begin{bmatrix} q_0 \\ q_1 \\ q_2 \\ q_3 \end{bmatrix} = \begin{bmatrix} \cos(\alpha/2) \\ e_1\sin(\alpha/2) \\ e_2\sin(\alpha/2) \\ e_3\sin(\alpha/2) \end{bmatrix}$$

where

$$q_0^2 + q_1^2 + q_2^2 + q_3^2 = 1$$

Relating Quaternions and the Coordinate Transformation Matrix. The direction cosine matrix in terms of Euler parameters is now

$$\underset{=}{C}_q = (q_0^2 - \underset{\sim}{q}^{\mathrm{T}}\underset{\sim}{q})\underset{=}{E} + 2\underline{q}\underline{q}^{\mathrm{T}} - 2q_0\underset{=}{S}(\underset{\sim}{q})$$

where $\underset{\sim}{q} = [q_1,\ q_2,\ q_3]^{\mathrm{T}}$, and $\underset{=}{E}$ is the identity matrix. The direction cosine matrix is now written in terms of quaternions

$$\underset{=}{C}_q = \begin{bmatrix} q_0^2 + q_1^2 - q_2^2 - q_3^2 & 2(q_1q_2 + q_3q_0) & 2(q_1q_3 - q_2q_4) \\ 2(q_1q_2 - q_3q_0) & q_0^2 - q_1^2 + q_2^2 - q_3^2 & 2(q_2q_3 + q_1q_4) \\ 2(q_1q_3 + q_2q_0) & 2(q_1q_2 + q_3q_0) & q_0^2 - q_1^2 - q_2^2 + q_3^2 \end{bmatrix}$$

It is possible to find the quaternions and the elements of the direction cosine matrix independently by integrating the angular rates about the principal axes of a body. Given the direction cosine matrix elements, we can find the quaternions, and vice versa. For a more extended discussion and application, the reader is referred to the listed references.

Dynamic Properties of a Rigid Body

Inertia Properties

The moments and products of inertia describe the distribution of mass for a body relative to a given coordinate system. This description relies on the specific orientation and reference frame. It is presumed that the reader is familiar with basic properties such as mass center, and the focus here is on those properties essential in understanding the general motion of rigid bodies, and particularly the rotational dynamics.

Moment of Inertia. For the rigid body shown in Fig. 9.33(a), the moment of inertia for a differential element, dm, about any of the three coordinate axes is defined as the product of the mass of the differential element and the square of the shortest distance from the axis to the element. As shown, $r_x = \sqrt{y^2 + z^2}$, so the contribution to the moment of inertia about the x-axis, I_{xx}, from dm is

$$dI_{xx} = r_x^2 = (y^2 + z^2)dm$$

The total I_{xx}, I_{yy}, and I_{zz} are found by integrating these expressions over the entire mass, m, of the body. In summary, the three moments of inertia about the x, y, and z axes are

$$\begin{aligned} I_{xx} &= \int_m r_x^2\,dm = \int_m (y^2 + z^2)\,dm \\ I_{yy} &= \int_m r_y^2\,dm = \int_m (x^2 + z^2)\,dm \\ I_{zz} &= \int_m r_z^2\,dm = \int_m (x^2 + y^2)\,dm \end{aligned} \tag{9.22}$$

Note that the moments of inertia, by virtue of their definition using squared distances and finite mass elements, are always positive quantities.

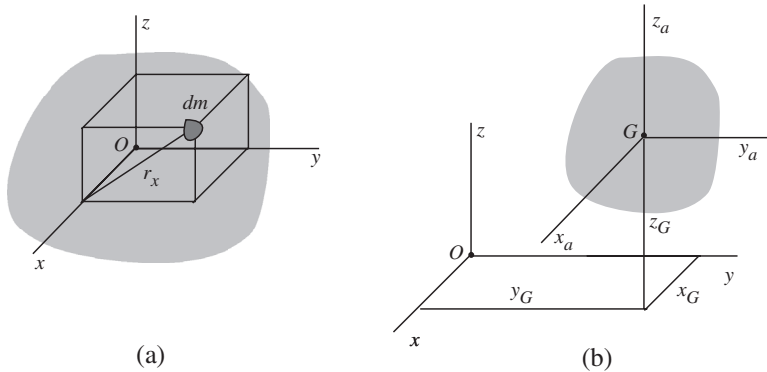



FIGURE 9.33 Rigid body properties are defined by how mass is distributed throughout the body relative to a specified coordinate system. (a) Rigid body used to describe moments and products of inertia. (b) Rigid body and axes used to describe parallel-axis and parallel-plane theorem.

Product of Inertia. The product of inertia for a differential element dm is defined with respect to a set of two orthogonal planes as the product of the mass of the element and the perpendicular (or shortest) distances from the planes to the element. So, with respect to the $y - z$ and $x - z$ planes (z common axis to these planes), the contribution from the differential element to I_{xy} is dI_{xy} and is given by $dI_{xy} = xy\,dm$.

As for the moments of inertia, by integrating over the entire mass of the body for each combination of planes, the products of inertia are

$$
\begin{aligned}
I_{xy} &= I_{yx} = \int_m xy\,dm \\
I_{yz} &= I_{zy} = \int_m yz\,dm \\
I_{xz} &= I_{zx} = \int_m xz\,dm
\end{aligned}
\tag{9.23}
$$

The product of inertia can be positive, negative, or zero, depending on the sign of the coordinates used to define the quantity. If either one or both of the orthogonal planes are planes of symmetry for the body, the product of inertia with respect to those planes will be zero. Basically, the mass elements would appear as pairs on each side of these planes.

Parallel-Axis and Parallel-Plane Theorems. The parallel-axis theorem can be used to transfer the moment of inertia of a body from an axis passing through its mass center to a parallel axis passing through some other point (see also the section "Kinetic Energy Storage"). Often the moments of inertia are known for axes fixed in the body, as shown in Fig. 9.33(b). If the center of gravity is defined by the coordinates (x_G, y_G, z_G) in the x, y, z axes, the parallel-axis theorem can be used to find moments of inertia relative to the x, y, z axes, given values based on the body-fixed axes. The relations are

$$
\begin{aligned}
I_{xx} &= (I_{xx})_a + m(y_G^2 + z_G^2) \\
I_{yy} &= (I_{yy})_a + m(x_G^2 + z_G^2) \\
I_{zz} &= (I_{zz})_a + m(x_G^2 + y_G^2)
\end{aligned}
$$

where, for example, $(I_{xx})_a$ is the moment of inertia relative to the x_a axis, which passes through the center of gravity. Transferring the products of inertia requires use of the parallel-plane theorem, which provides the relations

$$
\begin{aligned}
I_{xy} &= (I_{xy})_a + m x_G y_G \\
I_{yz} &= (I_{yz})_a + m y_G z_G \\
I_{zx} &= (I_{zx})_a + m z_G x_G
\end{aligned}
$$

Inertia Tensor. The rotational dynamics of a rigid body rely on knowledge of the inertial properties, which are completely characterized by nine terms of an inertia tensor, six of which are independent. The inertia tensor is

$$
\underline{I} = \begin{bmatrix} I_{xx} & -I_{xy} & -I_{xz} \\ -I_{yx} & I_{yy} & -I_{yz} \\ -I_{zx} & -I_{zy} & I_{zz} \end{bmatrix}
$$

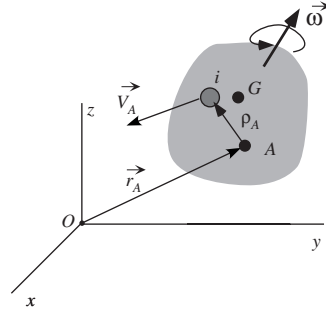



FIGURE 9.34 Rigid body in general motion relative to an inertial coordinate system, x, y, z.

and it relies on the specific location and orientation of coordinate axes in which it is defined. For a rigid body, an origin and axes orientation can be found for which the inertia tensor becomes diagonalized, or

$$\underline{I} = \begin{bmatrix} I_x & 0 & 0 \\ 0 & I_y & 0 \\ 0 & 0 & I_z \end{bmatrix}$$

The orientation for which this is true defines the principal axes of inertia, and the principal moments of inertia are now $I_x = I_{xx}$, $I_y = I_{yy}$, and $I_z = I_{zz}$ (one should be a maximum and another a minimum of the three). Sometimes this orientation can be determined by inspection. For example, if two of the three orthogonal planes are planes of symmetry, then all of the products of inertia are zero, so this would define principal axes of inertia.

The principal axes directions can be interpreted as an eigenvalue problem, and this allows you to find the orientation that will lead to principal directions, as well as define (transform) the inertia tensor into any orientation. For details on this method, see Crandall et al. [8].

Angular Momentum

For the rigid body shown in Fig. 9.34, conceptualized to be composed of particles, i, of mass, m_i, the angular momentum about the point A is defined as

$$(\vec{h}_A)_i = \vec{\rho}_A \times m_i \vec{V}_i$$

where $\vec{V}_i$ is the velocity measured relative to the inertial frame. Since $\vec{V}_i = \vec{V}_A + \vec{\omega} \times \vec{\rho}_A$, then

$$(\vec{h}_A)_i = \vec{\rho}_A \times m_i \vec{V}_i = m_i \vec{\rho}_A \times \vec{V}_A + m_i \vec{\rho}_A \times (\vec{\omega} \times \vec{\rho}_A)$$

Integrating over the mass of the body, the total angular momentum of the body is

$$\vec{h}_A = \left(\int_m \vec{\rho}_A dm\right) \times \vec{V}_A + \int_m \vec{\rho}_A \times (\vec{\omega} \times \vec{\rho}_A)\, dm \tag{9.24}$$

This equation can be used to find the angular momentum about a point of interest by setting the point A: (1) fixed, (2) at the center of mass, and (3) an arbitrary point on the mass. A general form arises in cases 1 and 2 that take the form

$$\vec{h} = \int_m \vec{\rho} \times (\vec{\omega} \times \vec{\rho})\, dm$$

When this form is expanded for either case into x, y, z components, then

$$\vec{h} = h_x\hat{i} + h_y\hat{j} + h_z\hat{k} = \int_m (x\hat{i} + y\hat{j} + z\hat{k}) \times [(\omega_x\hat{i} + \omega_y\hat{j} + \omega_z\hat{k}) \times (x\hat{i} + y\hat{j} + z\hat{k})]\, dm$$

which can be expanded to

$$\begin{aligned} h_x\hat{i} + h_y\hat{j} + h_z\hat{k} &= \left[\omega_x \int_m (y^2 + z^2)\, dm - \omega_y \int_m xy\, dm - \omega_z \int_m xz\, dm\right]\hat{i} \\ &= \left[-\omega_x \int_m xy\, dm + \omega_y \int_m (x^2 + z^2)\, dm - \omega_z \int_m yz\, dm\right]\hat{j} \\ &= \left[-\omega_x \int_m xy\, dm - \omega_y \int_m zy\, dm - \omega_z \int_m (x^2 + y^2)\, dm\right]\hat{k} \end{aligned}$$

The expression for moments and products of inertia can be identified here, and then this expression leads to the three angular momentum components, written in matrix form

$$\underline{h} = \begin{bmatrix} I_{xx} & -I_{xy} & -I_{xz} \\ -I_{yx} & I_{yy} & -I_{yz} \\ -I_{zx} & -I_{zy} & I_{zz} \end{bmatrix} \begin{bmatrix} \omega_x \\ \omega_y \\ \omega_z \end{bmatrix} = \bar{\underline{I}}\underline{\omega} \tag{9.25}$$

Note that the case where principal axes are defined leads to the much simplified expression

$$\vec{h} = I_{xx}\omega_x\hat{i} + I_{yy}\omega_y\hat{j} + I_{zz}\omega_z\hat{k}$$

This shows that when the body rotates so that its axis of rotation is parallel to a principal axis, the angular momentum vector, $\vec{h}$, is parallel to the angular velocity vector. In general, this is not true (this is related to the discussion at the end of the section “Inertia Properties”).

The angular momentum about an arbitrary point, Case 3, is the resultant of the angular momentum about the mass center (a free vector) and the moment of the *translational momentum* through the mass center,

$$\vec{p} = mV_x\hat{i} + mV_y\hat{j} + mV_z\hat{k} = m\vec{V}$$

or

$$\vec{h} = \vec{h}_G + \vec{r} \times \vec{p}$$

where $\vec{r}$ is the position vector from the arbitary point of interest to the mass center, G. This form can also be expanded into its component forms, as in Eq. (9.25).

Kinetic Energy of a Rigid Body

Several forms of the kinetic energy of a rigid body are presented in this section. From the standpoint of a bond graph formulation, where kinetic energy storage is represented by an **I** element, Eq. (9.25) demonstrates that the rigid body has at least three ports for rotational energy storage. Adding the three translational degrees of freedom, a rigid body can have up to six independent energy storage “ports.”

A 3-port **I** element can be used to represent the rotational kinetic energy for the case of rotation about a fixed point (no translation). The constitutive relation is simply Eq. (9.25). The kinetic energy is then

$$T = \frac{1}{2}\vec{\omega}\cdot\vec{h}$$

where $\vec{h}$ is the angular momentum with an inertia tensor defined about the fixed point. If the axes are aligned with principal axes, then

$$T = \frac{1}{2}I_x\omega_x^2 + \frac{1}{2}I_y\omega_y^2 + \frac{1}{2}I_z\omega_z^2$$

The total kinetic energy for a rigid body that can translate and rotate, with angular momentum defined with reference to the center of gravity, is given by

$$T = \frac{1}{2}mV_G^2 + \frac{1}{2}\vec{\omega}\cdot\vec{h}_G$$

where $V_G^2 = V_x^2 + V_y^2 + V_z^2$.

Rigid Body Dynamics

Given descriptions of inertial properties, translational and angular momentum, and kinetic energy of a rigid body, it is possible to describe the dynamics of a rigid body using the equations of motion using Newton's laws. The classical Euler equations are presented in this section, and these are used to show how a bond graph formulation can be used to integrate rigid body elements into a bond graph model.

Basic Equations of Motion

The translational momentum of the body in Fig. 9.30 is $\underline{p} = m\underline{V}$, where m is the mass, and $\underline{V}$ is the velocity of the mass center with three components of velocity relative to the inertial reference frame x_o, y_o, z_o. In three-dimensional motion, the net force on the body is related to the rate of change of momentum by Newton's law, namely,

$$\underline{F} = \frac{d}{dt}\underline{p}$$

which can be expressed as (using Eq. (9.9)),

$$\underline{F} = \left.\frac{\partial \underline{p}}{\partial t}\right|_{\text{rel}} + \underline{\Omega}\times\underline{p}$$

with $\underline{p}$ now relative to the moving frame x_a, y_a, z_a, and $\underline{\Omega}$ is the absolute angular velocity of the rotating axes.

A similar expression can be written for rate of change of the angular momentum, which is related to applied torques $\underline{T}$ by

$$\underline{T} = \left.\frac{\partial \underline{h}}{\partial t}\right|_{\text{rel}} + \underline{\Omega}\times\underline{h}$$

where $\underline{h}$ is relative to the moving frame x_a, y_a, z_a.

In order to use these relations effectively, the motion of the axes x_a, y_a, z_a, must be chosen to fit the problem at hand. This choice usually comes down to three cases described by how $\underline{\Omega}$ relates to the body angular velocity $\underline{\omega}$.

1. $\underline{\Omega} = 0$. If the body has general motion and the axes are chosen to translate with the center of mass, then this case will lead to a simple set of equations with $\Omega = 0$, although it will be necessary to describe the inertia properties of the body as functions of time.
2. $\underline{\Omega} \neq 0 \neq \underline{\omega}$. In this case, axes have an angular velocity different from that of the body, a form convenient for bodies that are symmetrical about their spinning axes. The moments and products of inertia will be constant relative to the rotating axes. The equations become

$$\begin{aligned}
F_x &= m\dot{V}_x - mV_y\Omega_z + mV_z\Omega_y \\
F_y &= m\dot{V}_y - mV_z\Omega_x + mV_x\Omega_z \\
F_z &= m\dot{V}_z - mV_x\Omega_y + mV_y\Omega_x \\
T_x &= I_x\dot{\omega}_x - I_y\omega_y\Omega_z + I_z\Omega_y\omega_z \\
T_y &= I_y\dot{\omega}_y - I_z\omega_z\Omega_x + I_x\Omega_z\omega_x \\
T_z &= I_z\dot{\omega}_z - I_x\omega_x\Omega_y + I_y\Omega_x\omega_y
\end{aligned} \tag{9.26}$$

3. $\underline{\Omega} = \underline{\omega}$. Here the axes are fixed and moving with the body. The moments and products of intertia relative to the moving axes will be constant. A particularly convenient case arises if the axes are chosen to be the principal axes of inertia (see the section titled "Inertia Properties"), which leads to the *Euler equations*,[4]

$$\begin{aligned}
F_x &= m\dot{V}_x - mV_y\omega_z + mV_z\omega_y \\
F_y &= m\dot{V}_y - mV_z\omega_x + mV_x\omega_z \\
F_z &= m\dot{V}_z - mV_x\omega_y + mV_y\omega_x \\
T_x &= I_x\dot{\omega}_x - (I_y - I_z)\omega_y\omega_z \\
T_y &= I_y\dot{\omega}_y - (I_z - I_x)\omega_z\omega_x \\
T_z &= I_z\dot{\omega}_z - (I_x - I_y)\omega_x\omega_y
\end{aligned} \tag{9.27}$$

These equations of motion can be used to determine the forces and torques, given motion of the body. Textbooks on dynamics [12,23] provide extensive examples on this type of analysis. Alternatively, these can be seen as six nonlinear, coupled ordinary differential equations (ODEs). Case 3 (the Euler equations) could be solved in such a case, since these can be rewritten as six first-order ODEs. A numerical solution may need to be implemented. Modern computational software packages will readily handle these equations, and some will feature a form of these equations in a form suitable for immediate use. Case 2 requires knowledge of the axes' angular velocity, $\underline{\Omega}$.

If the rotational motion is coupled to the translational motion such that the forces and torques, say, are related, then a dynamic model is required. In some, it may be desirable to formulate the problem in a bond graph form, especially if there are actuators and sensors and other multienergetic systems to be incorporated.

---

[4]First developed by the Swiss mathematician L. Euler.

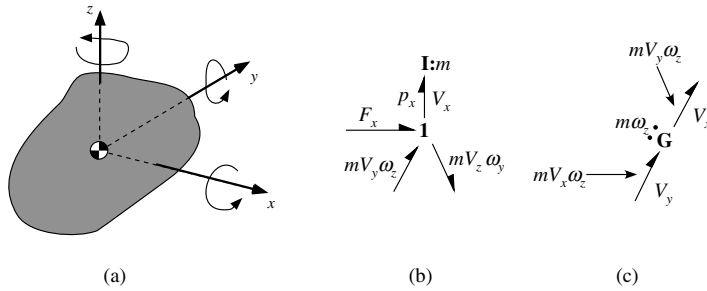



FIGURE 9.35 (a) Rigid body with angular velocity components about x, y, z axes. (b) x-direction translational dynamics in bond graph form. (c) Gyrator realization of coupling forces.

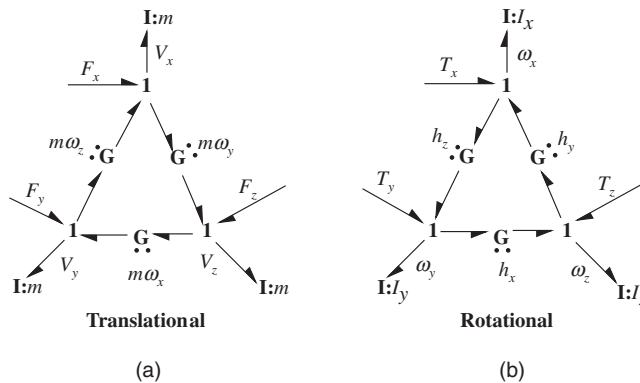



FIGURE 9.36 (a) Bond graph for rigid body translation. (b) Bond graph for rigid body rotation.

Rigid Body Bond Graph Formulation

Due to the body's rotation, there is an inherent coupling of the translational and rotational motion, which can be summarized in a bond graph form. Consider the case of Euler's equations, given in Eqs. (9.27). For the x-direction translational dynamics,

$$F_x = \dot{p}_x - mV_y\omega_z + mV_z\omega_y$$

where $p_x = mV_x$, and F_x is the net "external" applied forces in the x-direction. This equation, a summation of forces (efforts) is represented in bond graph form in Fig. 9.35(b). All of these forces are applied at a common velocity, V_x, represented by the 1-junction. The **I** element represents the storage of kinetic energy in the body associated with motion in the x-direction. The force $mV_y\omega_z$ in Fig. 9.35(b) is induced by the y-direction velocity, V_y, and by the angular velocity component, ω_z. This physical effect is gyrational in nature, and can be captured by the gyrator, as shown in Fig. 9.35(c). Note that this is a modulated gyrator (could also be shown as **MGY**) with a gyrator modulus of $r = m\omega_z$ (verify that the units are force).

The six equations of motion, Eqs. (9.27), can be represented in bond graph form as shown in Fig. 9.36. Note that these two bond graph ring formations, first shown by Karnopp and Rosenberg [18], capture the Euler equations very efficiently and provide a graphical mnemonic for rigid body motion. Indeed, Euler's equations can now be "drawn" simply in the following steps: (1) lay down three 1-junctions representing angular velocity about x, y, z (counter clockwise labeling), with **I** elements attached, (2) between each 1-junction place a gyrator, modulated by the momentum about the axis represented by the

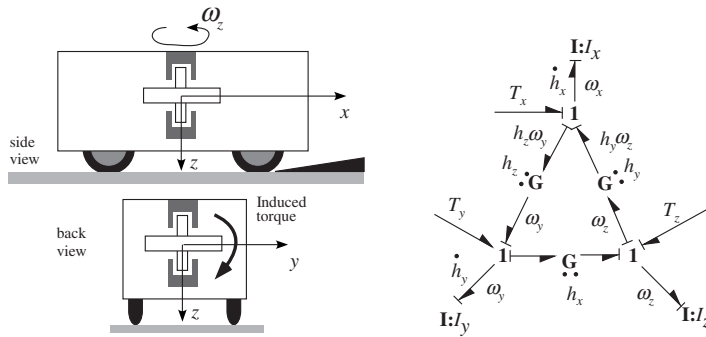



FIGURE 9.37 A cart with a rigid and internally mounted flywheel approaches a ramp.

1-junction directly opposite in the triangle, (3) draw power arrows in a counter clockwise direction. This sketch will provide the conventional Euler equations. The translational equations are also easily sketched.

These bond graph models illustrate the inherent coupling through the gyrator modulation. There are six **I** elements, and each can represent an independent energetic state in the form of the momenta [p_x, p_y, p_z, h_x, h_y, h_z] or alternatively the analyst could focus on the associated velocities [V_x, V_y, V_z, ω_x, ω_y, ω_z].

If forces and torques are considered as inputs, through the indicated bonds representing F_x, F_y, F_z, T_x, T_y, T_z, then you can show that all the **I** elements are in integral causality, and the body will have six independent states described by six first-order nonlinear differential equations.

Example: Cart-Flywheel

A good example of how the rigid body bond graphs represent the basic mechanics inherent to Eqs. (9.27) and of how the graphical modeling can be used for "intuitive" gain is shown in Fig. 9.37. The flywheel is mounted in the cart, and spins in the direction shown. The body-fixed axes are mounted in the vehicle, with the convention that z is positive into the ground (common in vehicle dynamics). The cart approaches a ramp, and the questions which arise are whether any significant loads will be applied, what their sense will be, and on which parameters or variables they are dependent.

The bond graph for rotational motion of the flywheel (assume it dominates the problem for this example) is shown in Fig. 9.37. If the flywheel momentum is assumed very large, then we might just focus on its effect. At the 1-junction for ω_x, let $T_x = 0$, and since ω_z is spinning in a negative direction, you can see that the torque $h_z\omega_y$ is applied in a positive direction about the x-axis. This will tend to "roll" the vehicle to the right, and the wheels would feel an increased normal load. With the model shown, it would not be difficult to develop a full set of differential equations.

Need for Coordinate Transformations

In the cart-flywheel example, it is assumed that as the front wheels of the cart lift onto the ramp, the flywheel will react because of the direct induced motion at the bearings. Indeed, the flywheel-induced torque is also transmitted directly to the cart. The equations and basic bond graphs developed above are convenient if the forces and torques applied to the rigid body are moving with the rotating axes (assumed to be fixed to the body). The orientational changes, however, usually imply that there is a need to relate the body-fixed coordinate frames or axes to inertial coordinates. This is accomplished with a coordinate transformation, which relates the body orientation into a frame that makes it easier to interpret the motion, apply forces, understand and apply measurements, and apply feedback controls.

Example: Torquewhirl Dynamics

Figure 9.38(a) illustrates a cantilevered rotor that can exhibit torquewhirl. This is a good example for illustrating the need for coordinate transformations, and how Euler angles can be used in the modeling process. The whirling mode is conical and described by the angle θ. There is a drive torque, T_s, that is

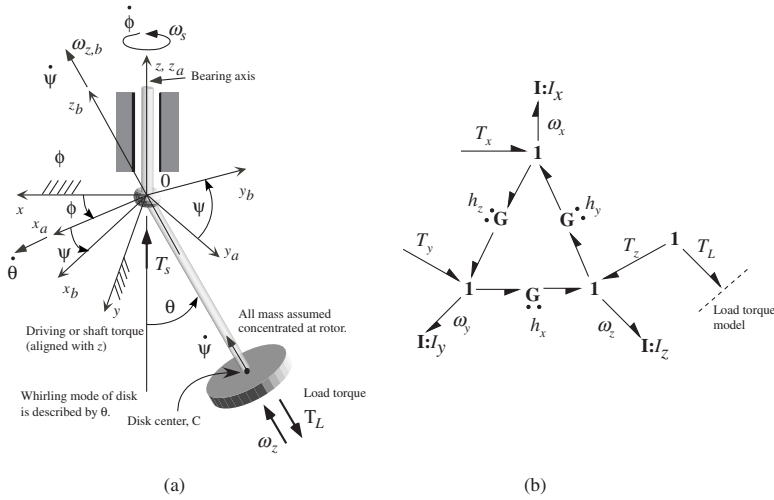



FIGURE 9.38 (a) Cantilevered rotor with flexible joint and rigid shaft (after Vance [36]). (b) Bond graph representing rigid body rotation of rotor.

aligned with the bearing axis, z, where x, y, z is the inertial coordinate frame. The bond graph in Fig. 9.38(b) captures the rigid body motion of the rotor, represented in body-fixed axes x_b, y_b, z_b, which represent principal axes of the rotor.

The first problem seen here is that while the bond graph leads to a very convenient model formulation, the applied torque, T_s, is given relative to the inertial frame x, y, z. Also, it would be nice to know how the rotor moves relative to the inertial frame, since it is that motion that is relevant. Other issues arise, including a stiffness of the rotor that is known relative to the angle θ. These problems motivate the use of Euler angles, which will relate the motion in the body fixed to the inertial frame, and provide three additional state equations for ϕ, θ, and ψ (which are needed to quantify the motion).

In this example, the rotation sequence is (1) x, y, z (inertial) to x_a, y_b, z_c, with ϕ about the z-axis, so note, $\dot{\phi} = \omega_s$, (2) x_a, y_a, z_a to x_b, y_b, z_b, with θ about x_a, (3) ψ rotation about z_b. Our main interest is in the overall transformation from x, y, z (inertia) to x_b, y_b, z_b (body-fixed). In this way, we relate the body angular velocities to inertial velocities using the relation from Eq. (9.20),

$$\begin{bmatrix} \omega_x \\ \omega_y \\ \omega_z \end{bmatrix}_b = \begin{bmatrix} \dot{\phi}\sin\theta\sin\psi + \dot{\theta}\cos\psi \\ \dot{\phi}\sin\theta\cos\psi - \dot{\theta}\sin\psi \\ \dot{\phi}\cos\theta + \dot{\psi} \end{bmatrix}$$

where the subscript b on the left-hand side denotes velocities relative to the x_b, y_b, z_b axes. A full and complete bond graph would include a representation of these transformations (e.g., see Karnopp, Margolis, and Rosenberg [17]). Explicit 1-junctions can be used to identify velocity junctions at which torques and forces are applied. For example, at a 1-junction for $\dot{\phi} = \omega_z$, the input torque T_s is properly applied. Once the bond graph is complete, causality is applied. The preferred assignment that will lead to integral causality on all the **I** elements is to have torques and forces applied as causal *inputs*. Note that in transforming the expression above which relates the angular velocities, a problem with Euler angles arises related to the singularity (here at $\theta = \pi/2$, for example).

An alternative way to proceed in the analysis is using a Lagrangian approach as in Section 9.7, as done by Vance [36] (see p. 292). Also, for advanced multibody systems, a multibond formulation can be more efficient and may provide insight into complex problems (see Breedveld [4] or Tiernego and Bos [35]).

9.7 Lagrange's Equations

The discussion on energy methods focuses on deriving constitutive relations for energy-storing multiports, and this can be very useful in some modeling exercises. For some cases where the constraint relationships between elements are primarily holonomic, and definitely scleronomic (not an explicit function of time), implicit multiport fields can be formulated (see Chapter 7 of [17]). The principal concern arises because of dependent energy storage, and the methods presented can be a solution in some practical cases. However, there are many mechanical systems in which geometric configuration complicates the matter. In this section, Lagrange's equations are introduced to facilitate analysis of those systems.

There are several ways to introduce, derive, and utilize the concepts and methods of Lagrange's equations. The summary presented below is provided in order to introduce fundamental concepts, and a thorough derivation can be found either in Lanczos [20] or Goldstein [11]. A derivation using energy and power flow is presented by Beaman, Paynter, and Longoria [3].

Lagrange's equations are also important because they provide a unified way to model systems from different energy domains, just like a bond graph approach. The use of scalar energy functions and minimal geometric reasoning is preferred by some analysts. It is shown in the following that the particular benefits of a Lagrange approach that make it especially useful for modeling mechanical systems enhance the bond graph approach. A combined approach exploits the benefits of both methods, and provides a methodology for treating complex mechatronic systems in a systematic fashion.

Classical Approach

A classical derivation of Lagrange's equations evolves from the concept of virtual displacement and virtual work developed for analyzing static systems (see Goldstein [11]). To begin with, the Lagrange equations can be derived for dynamic systems by using Hamilton's principle or D'Alembert's principle.

For example, for a system of particles, Newton's second law for the i mass, $\mathbf{F}_i = \mathbf{p}_i$, is rewritten, $\mathbf{F}_i - \mathbf{p}_i = 0$. The forces are classified as either applied or constraint, $\mathbf{F}_i = \mathbf{F}_i^{(a)} + \mathbf{f}_i$. The principle of virtual work is applied over the system, recognizing that constraint forces $\mathbf{f}_i$, do no work and will drop out. This leads to the D'Alembert principle [11],

$$\sum_i (\mathbf{F}_i^{(a)} - \dot{\mathbf{p}}_i) \cdot \delta \mathbf{r}_i = 0 \tag{9.28}$$

The main point in presenting this relation is to show that: (a) the constraint forces do not appear in this formulative equation and (b) the need arises for transforming relationships between, in this case, the N coordinates of the particles, $\mathbf{r}_i$, and a set of n **generalized coordinates**, $\mathbf{q}_i$, which are independent of each other (for holonomic constraints), i.e.,

$$\mathbf{r}_i = \mathbf{r}_i(q_1, q_2, \ldots, q_n, t) \tag{9.29}$$

By transforming to generalized coordinates, D'Alembert's principle becomes [11]

$$\sum_j \left[\left\{ \frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_j}\right) - \frac{\partial T}{\partial q_j} \right\} - Q_j \right] \delta q_j = 0 \tag{9.30}$$

where T is the system kinetic energy, and the Q_j are components of the **generalized forces** given by

$$Q_j = \sum_i \mathbf{F}_i \cdot \frac{\partial \mathbf{r}_i}{\partial q_j}$$

If the transforming relations are restricted to be holonomic, the constraint conditions are implicit in the transforming relations, and independent coordinates are assured. Consequently, all the terms in Eq. (9.30) must vanish for independent virtual displacements, δq_j, resulting in the n equations:

$$\frac{d}{dt}\left(\frac{\partial T}{\partial \dot{q}_j}\right) - \frac{\partial T}{\partial q_j} = Q_j \tag{9.31}$$

These equations become Lagrange's equations through the following development. Restrict all the applied forces, Q_j, to be derivable from a scalar function, U, where in general, $U = U(q_j, \dot{q}_j)$, and

$$Q_j = -\frac{\partial U}{\partial q_j} + \frac{d}{dt}\left(\frac{\partial U}{\partial \dot{q}_j}\right)$$

The Lagrangian is defined as $L = T - U$, and substituted into Eq. (9.31) to yield the n Lagrange equations:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_j}\right) - \frac{\partial L}{\partial q_j} = Q_j \tag{9.32}$$

This formulation yields n second-order ODEs in the q_j.

Dealing with Nonconservative Effects

The derivation of Lagrange's equations assumes, to some extent, that the system is conservative, meaning that the total of kinetic and potential energy remains constant. This is not a limiting assumption because the process of reticulation provides a way to extract nonconservative effects (inputs, dissipation), and then to assemble the system later. It is necessary to recognize that the nonconservative effects can be integrated into a model based on Lagrange's equations using the Q_i's. Associating these forces with the generalized coordinates implies work is done, and this is in accord with energy conservation principles (we account for total work done on system). The generalized force associated with a coordinate, q_i, and due to external forces is then derived from $Q_i = \delta W_i / \delta q_i$, where W_i is the work done on the system by all external forces during the displacement, δq_i.

Extensions for Nonholonomic Systems

In the case of nonholonomic constraints, the coordinates q_j are not independent. Assume you have m nonholonomic constraints ($m \leq n$). If the equations of constaint can be put in the form

$$\sum_k \frac{\partial a_l}{\partial q_k} dq_k + \frac{\partial a_l}{\partial t} dt = \sum_k a_{lk} dq_k + a_{lt}\, dt = 0 \tag{9.33}$$

where l indexes up to m such constraints, then the Lagrange equations are formulated with Lagrange undetermined multipliers, λ_l. We maintain n coordinates, q_k, but the n Lagrange equations are now expressed [11] as

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right) - \frac{\partial L}{\partial q_k} = \sum_l \lambda_l a_{lk}, \quad k = 1, 2, \ldots, n \tag{9.34}$$

However, since there are now m unknown Lagrange multipliers, λ_l, it is necessary to solve an additional m equations:

$$\sum_k a_{lk} \dot{q}_k + a_{lt} = 0 \tag{9.35}$$

The terms $\sum_l \lambda_l a_{lk}$ can be interpreted as generalized forces of constraint. These are still workless constraints. The Lagrange equations for nonholonomic constraints can be used to study holonomic systems, and this analysis would provide a solution for the constraint forces through evaluation of the Lagrange multipliers. The use of Lagrange's equations with Lagrange multipliers is one way to model complex, constrained multibody systems, as discussed in Haug [14].

Mechanical Subsystem Models Using Lagrange Methods

The previous sections summarize a classical formulation and application of Lagrange's equations. When formulating models of mechanical systems, these methods are well proven. Lagrange's equations are recognized as an approach useful in handling systems with complex mechanical systems, including systems with constraints. The energy-basis also makes the method attractive from the standpoint of building multi-energetic system models, and Lagrange's equations have been used extensively in electromechanics modeling, for example. For conservative systems, it is possible to arrive at solutions sometimes without worrying about forces, especially since nonconservative effects can be handled "outside" the conservative dynamics. Developing transformation equations between the coordinates, say $\mathbf{x}$, used to describe the system and the independent coordinates, $\mathbf{q}$, helps assure a minimal formulation. However, it is possible sometimes to lose insight into cause and effect, which is more evident in other approaches. Also, the algebraic burden can become excessive. However, it is the analytical basis of the method that makes it especially attactive. Indeed, with computer-aided symbolic processing techniques, extensive algebra becomes a non-issue.

In this section, the advantages of the Lagrange approach are merged with those of a bond graph approach. The concepts and formulations are classical in nature; however, the graphical interpretation adds to the insight provided. Further, the use of bond graphs assures a consistent formulation with causality so that some insight is provided into how the conservative dynamics described by the energy functions depend on inputs, which typically arrive from the nonconservative dynamics. The latter are very effectively dealt with using bond graph methods, and the combined approach is systematic and yields first-order differential equations, rather than the second-order ODEs in the classical approach. Also, it will be shown that in some cases the combined approach makes it relatively easy to model certain systems that would be very troublesome for a direct approach by either method independently.

A Lagrange bond graph subsystem model will capture the elements summarized with a word bond graph in Fig. 9.39. The key elements are identified as follows: (a) conservative energy storage captured by kinetic and potential energy functions, (b) power-conserving transforming relations, and (c) coupling/interconnections with nonconservative and non-Lagrange system elements. Note that on the nonconservative side of the transforming relations, there are m coordinates that can be identified in the modeling, but these are not independent. The power-conserving transforming relations reduce the coordinates to a set of n independent coordinates, q_i. Associated with each independent coordinate or velocity, $\dot{q}_i$, there is an associated storage of kinetic and potential energy which can be represented by the coupled **IC** in Fig. 9.40(a) [16]. An alternative is the single **C** element used to capture all the coupled energy storage [3], where the gyrator has a modulus of 1 (this is called a symplectic gyrator). In either case, this structure shows that there will be one common flow junction associated with each independent coordinate. Recall the efforts at a 1-junction sum, and at this ith junction,

$$E_{q_i} = \dot{\tilde{p}}_i + e_{q_i} \tag{9.36}$$

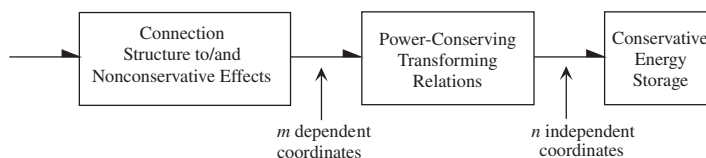



FIGURE 9.39 Block diagram illustrating the Lagrange subsystem model.

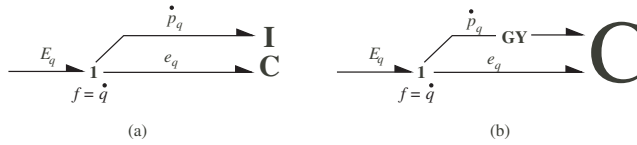



FIGURE 9.40 Elementary formulation of a flow junction in a Lagrange subsystem model. The efforts at the 1-junction for this ith independent flow variable, $\dot{q}_i$, represent Lagrange's equations.

where E_{q_i} is the net nonconservative effort at $\dot{q}_i$, e_{q_i} is a generalized conservative effort that will be determined by the Lagrange system, and the effort $\tilde{\dot{p}}_i$ is a rate of change of an ith generalized momentum. These terms will be defined in the next section. However, note that this effort sum is simply Newton's laws derived by virtue of a Lagrange formulation. In fact, this equation is simply a restatement of the ith Lagrange equation, as will be shown in the following. These effort sum equations give n first-order ODEs by solving for $\dot{p}_i$. The other n equations will be for the displacement variables, q_i. The following methodology is adapted from Beaman, Paynter, and Longoria [3].

Methodology for Building Subsystem Model

Conduct Initial Modeling. Isolate the conservative parts of the system, and make sure that any constraints are holonomic. This reticulation will identify ports to the system under study, including points in the system (typically velocities) where forces and/or torques of interest can be applied (e.g., at flow junctions). These forces and torques are either nonconservative, or they are determined by a system external to the Lagrange-type subsystem. This is a modeling decision. For example, a force due to gravity could be included in a Lagrange subsystem (being conservative) or it could be shown explicity at a velocity junction corresponding to motion modeled outside of the Lagrange subsystem. This will be illustrated in one of the examples that follow.

Define Generalized Displacement Variables. In a Lagrange approach, it is necessary to identify variables that define the configuration of a system. In mechanical system, these are translational and rotational displacements. Further, these variables are typically associated with the motion or relative motion of bodies. To facilitate a model with a minimum and independent set of coordinates, develop transforming relations between the m velocities or, more generally, flows $\dot{\mathbf{x}}$, and n independent flows, $\dot{\mathbf{q}}$. The form is [3],

$$\dot{\mathbf{x}} = \mathbf{T}(\mathbf{q})\dot{\mathbf{q}} \tag{9.37}$$

explicity showing that the matrix $\mathbf{T}(\mathbf{q})$ can depend on $\mathbf{q}$. This can be interpreted, in bond graph modeling terms, as a modulated transformer relationship, where $\mathbf{q}$ contains the modulating variables. The independent generalized displacements, $\mathbf{q}$, will form possible state variables of the Lagrange subsystem.

The transforming relationships are commonly derived from (holonomic) constraints, and from considerations of geometry and basic kinematics. The matrix $\mathbf{T}$ is $m \times n$ and may not be invertible. The bond graph representation is shown in Fig. 9.41.

Formulate the Kinetic Energy Function. Given the transforming relationships, it is now possible to express the total kinetic energy of the Lagrange subsystem using the independent flow variables, $\dot{\mathbf{q}}$. First, the kinetic energy can be written using the $\dot{\mathbf{x}}$ (this is usually easier), or $\mathbf{T} = \mathbf{T}_{\dot{\mathbf{x}}}(\dot{\mathbf{x}})$. Then the relations in Eq. (9.37) are used to transform this kinetic energy function so it is expressed as a function of the $\mathbf{q}$ and $\dot{\mathbf{q}}$ variables, $\mathbf{T}_{\dot{\mathbf{x}}}(\dot{\mathbf{x}}) \rightarrow \mathbf{T}_{\dot{\mathbf{q}}\mathbf{q}}(\dot{\mathbf{q}}, \mathbf{q})$. For brevity, this can be indicated in the subscript, or just $\mathbf{T}_{\dot{\mathbf{q}}\mathbf{q}}$. For example, a kinetic energy function that depends on x, θ, and $\dot{\theta}$ is referred to as $T_{\dot{\theta}\theta x}$ (if the number of variables is very high, certainly such a convention would not be followed).

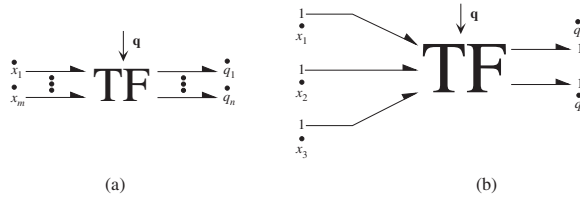



FIGURE 9.41 (a) Bond graph representation of the transforming relations. (b) Example for the case where $m = 3$ and $n = 2$.

Define Generalized Momentum Variables. With the kinetic energy function now in terms of the independent flows, $\dot{\mathbf{q}}$, generalized momenta can be defined as [3,20],

$$\tilde{\mathbf{p}} = \frac{\partial \mathbf{T}_{\dot{q}q}}{\partial \dot{\mathbf{q}}} \tag{9.38}$$

where the "tilde" ($\tilde{\mathbf{p}}$) notation is used to distinguish these momentum variables from momentum variables defined strictly through the principles summarized in Table 9.5. In particular note that these generalized momentum variables may be functions of flow as well as of displacement (i.e., they may be configuration dependent).

Formulate the Potential Energy Function. In general, a candidate system for study by a Lagrange approach will store potential energy, in addition to kinetic energy, and the potential energy function, U, should be expressed in terms of the dependent variables, $\mathbf{x}$. Using the tranforming relations in Eq. (9.37), the expression is then a function of $\mathbf{q}$, or $U = U(\mathbf{q}) = U_{\mathbf{q}}$. In mechanical systems, this function is usually formed by considering energy stored in compliant members, or energy stored due to a gravitational potential. In these cases, it is usually possible to express the potential energy function in terms of the displacement variables, $\mathbf{q}$.

Derive Generalized Conservative Efforts. A conservative effort results and can be found from the expression

$$\tilde{\mathbf{e}}_{\mathbf{q}} = -\frac{\partial \mathbf{T}_{\dot{q}q}}{\partial \mathbf{q}} + \frac{\partial U_{\mathbf{q}}}{\partial \mathbf{q}} \tag{9.39}$$

where the $\mathbf{q}$ subscript is used to denote these as conservative efforts. The first term on the right-hand side represents an effect due to dependence of kinetic energy on displacement, and the second term will be recognized as the potential energy derived effort.

Identify and Express Net Power Flow into Lagrange Subsystem. At the input to the Lagrange subsystem on the "nonconservative" side, the power input can be expressed in terms of effort and flow products. Since the transforming relations are power-conserving, this power flow must equal the power flow on the "conservative" side. This fact is expressed by

$$P_{\mathbf{x}} = \underbrace{\mathbf{e}_{\mathbf{x}}}_{1\times m}\ \underbrace{\dot{\mathbf{x}}}_{m\times 1} = \underbrace{\mathbf{e}_{\mathbf{x}}}_{1\times m}\ \underbrace{\mathbf{T}(\mathbf{q})}_{m\times n}\ \underbrace{\dot{\mathbf{q}}}_{n\times 1} = \underbrace{\mathbf{E}_{\mathbf{q}}}_{1\times n}\ \underbrace{\dot{\mathbf{q}}}_{n\times 1} \tag{9.40}$$

where the term $\mathbf{E}_{\mathbf{q}}$ is the nonconservative effort transformed into the $\mathbf{q}$ coordinates. This term can be computed as shown by

$$\mathbf{E}_{\mathbf{q}} = \mathbf{e}_{\mathbf{x}}\mathbf{T}(\mathbf{q}) \tag{9.41}$$

Summary of the Method. In summary, all the terms for a Lagrange subsystem can be systematically derived. There are some difficulties that can arise. To begin with, the first step can require some geometric reasoning, and often this can be a problem in some cases, although not insurmountable. The n

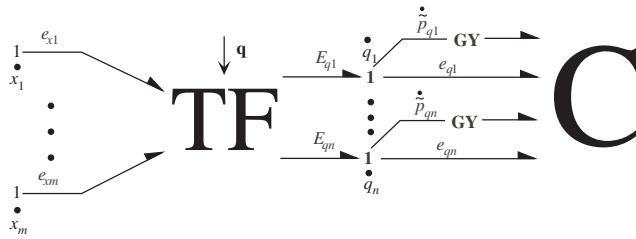



FIGURE 9.42 Lagrange subsystem model.

momentum state equations for this Lagrange subsystem are given by

$$\dot{\tilde{p}} = -e_i + E_i \tag{9.42}$$

and the state equations for the q_i must be found by inverting the generalized momentum equations, (9.38). In some cases, these n equations are coupled and must be solved simultaneously. In the end, there are $2n$ first-order state equations. In addition, the final bond graph element shown in Fig. 9.42 can be coupled to other systems to build a complex system model.

Note that in order to have the $2n$ equations in integral causality, efforts (forces and torques) should be specified as causal inputs to the transforming relations. Also, this subsystem model assumes that only holonomic constraints are applied. While this might seem restrictive, it turns out that, for many practical cases, the physical effects that lead to nonholonomic constraints can be dealt with "outside" of the Lagrange model, along with dissipative effects, actuators, and so on.

References

1. Arczewski, K. and Pietrucha, J., *Mathematical Modelling of Complex Mechanical Systems,* Ellis Horwood, New York, 1993.
2. Beaman, J.J. and Rosenberg, R.C., "Constitutive and modulation structure," *Journal of Dynamic Systems, Measurement, and Control (ASME),* Vol. 110, No. 4, pp. 395–402, 1988.
3. Beaman, J.J., Paynter, H.M., and Longoria, R.G., *Modeling of Physical Systems,* Cambridge University Press, in progress.
4. Breedveld, P.C., "Multibond graph elements in physical systems theory," *Journal of the Franklin Institute,* Vol. 319, No. 1–2, pp. 1–36, 1985.
5. Bedford, A. and Fowler, W., *Engineering Mechanics. Dynamics,* 2nd edition, Addison Wesley Longman, Menlo Park, CA, 1999.
6. Burr, A.H., *Mechanical Analysis and Design,* Elsevier Science Publishing, Co., New York, 1981.
7. Chou, J.C.K, "Quaternion kinematic and dynamic differential equations," *IEEE Transactions on Robotics and Automation,* Vol. 8, No. 1, February, 1992.
8. Crandall, S., Karnopp, D.C., Kurtz, E.F., and Pridmore-Brown, D.C., *Dynamics of Mechanical and Electromechanical Systems,* McGraw-Hill, New York, 1968 (Reprinted by Krieger Publishing Co., Malabar, FL, 1982).
9. Den Hartog, J.P., *Advanced Strength of Materials,* McGraw-Hill, New York, 1952.
10. Fjellstad, O. and Fossen, T.I., "Position and attitude tracking of AUVs: a quaternion feedback approach," *IEEE Journal of Oceanic Engineering,* Vol. 19, No. 4, pp. 512–518, 1994.
11. Goldstein, D., *Classical Mechanics,* 2nd edition, Addison-Wesley, Reading, MA, 1980.
12. Greenwood, D.T., *Principles of Dynamics,* Prentice-Hall, Englewood Cliffs, NJ, 1965.
13. Harding, C.F., "Solution to Euler's gyrodynamics-I," *Journal of Applied Mechanics,* Vol. 31, pp. 325–328, 1964.

14. Haug, E.J., *Computer Aided Kinematics and Dynamics of Mechanical Systems,* Allyn and Bacon, Needham, MA, 1989.
15. Kane, T.R. and Levinson, D.A., *Dynamics: Theory and Applications,* McGraw-Hill Publishing Co., New York, 1985.
16. Karnopp, D., "An approach to derivative causality in bond graph models of mechanical systems," *Journal of the Franklin Institute,* Vol. 329, No. 1, pp. 65–75, 1992.
17. Karnopp, D.C., Margolis, D., and Rosenberg, R.C., *System Dynamics: Modeling and Simulation of Mechatronic Systems,* Wiley, New York, 2000, 3rd edition, or *System Dynamics: A Unified Approach,* 1990, 2nd edition.
18. Karnopp, D. and Rosenberg, R.C., *Analysis and Simulation of Multiport Systems. The Bond Graph Approach to Physical System Dynamics,* MIT Press, Cambridge, MA, 1968.
19. Kuipers, J.B., *Quaternions and Rotation Sequences,* Princeton University Press, Princeton, NJ, 1998.
20. Lanczos, C., *The Variational Principles of Mechanics,* 4th edition, University of Toronto Press, Toronto, 1970. Also published by Dover, New York, 1986.
21. Lyshevski, S.E., *Electromechanical Systems, Electric Machines, and Applied Mechatronics,* CRC Press, Boca Raton, FL, 2000.
22. Matschinsky, W., *Road Vehicle Suspensions,* Professional Engineering Publishing Ltd., Suffolk, UK, 1999.
23. Meriam, J.L. and Kraige, L.G., *Engineering Mechanics. Dynamics,* 4th edition, John Wiley and Sons, New York, 1997.
24. Mortensen, R.E., "A globally stable linear regulator," *International Journal of Control,* Vol. 8, No. 3, pp. 297–302, 1968.
25. Nikravesh, P.E. and Chung, I.S., "Application of Euler parameters to the dynamic analysis of three-dimensional constrained mechanical systems," *Journal of Mechanical Design (ASME),* Vol. 104, pp. 785–791, 1982.
26. Nikravesh, P.E., Wehage, R.A., and Kwon, O.K., "Euler parameters in computational kinematics and dynamics, Parts 1 and 2," *Journal of Mechanisms, Transmissions, and Automation in Design (ASME),* Vol. 107, pp. 358–369, 1985.
27. Nososelov, V.S., "An example of a nonholonomic, nonlinear system not of the Chetaev type," *Vestnik Leningradskogo Universiteta,* No. 19, 1957.
28. Paynter, H., *Analysis and Design of Engineering Systems,* MIT Press, Cambridge, MA, 1961.
29. Roark, R.J. and Young, W.C., *Formulas for Stress and Strain,* McGraw-Hill, New York, 1975.
30. Roberson, R.E. and Schwertassek, *Dynamics of Multibody Systems,* Springer-Verlag, Berlin, 1988.
31. Rosenberg, R.M., *Analytical Dynamics of Discrete Systems,* Plenum Press, New York, 1977.
32. Rosenberg, R. and Karnopp, D., *Introduction to Physical System Dynamics,* McGraw-Hill, New York, 1983.
33. Rowell, D. and Wormley, D.N., *System Dynamics,* Prentice-Hall, Upper Saddle River, NJ, 1997.
34. Siciliano, B. and Villani, L., *Robot Force Control,* Kluwer Academic Publishers, Norwell, MA, 1999.
35. Tiernego, M.J.L. and Bos, A.M., "Modelling the dynamics and kinematics of mechanical systems with multibond graphs," *Journal of the Franklin Institute,* Vol. 319, No. 1–2, pp. 37–50, 1985.
36. Vance, J.M., *Rotordynamics of Turbomachinery,* John Wiley and Sons, New York, 1988.
37. Wehage, R.A., "Quaternions and Euler parameters—a brief exposition," in *Proceedings of the NATO Advanced Study Institute on Computer Aided Analysis and Optimization of Mechanical System Dynamics,* E.J. Haug (ed.), Iowa City, IA, August 1–12, 1983, pp. 147–182.
38. Wie, B. and Barba, P.M., "Quaternion feedback for spacecraft large angle maneuvers," *Journal of Guidance, Control, and Dynamics,* Vol. 8, pp. 360–365, May–June 1985.
39. Wittenburg, J., *Dynamics of Systems of Rigid Bodies,* B.G. Teubner, Studttgart, 1977.

10

Fluid Power Systems

Qin Zhang
University of Illinois

Carroll E. Goering
University of Illinois

10.1 Introduction

Fluid Power Systems

A fluid power system uses either liquid or gas to perform desired tasks. Operation of both the liquid systems (hydraulic systems) and the gas systems (pneumatic systems) is based on the same principles. For brevity, we will focus on hydraulic systems only.

A fluid power system typically consists of a hydraulic pump, a line relief valve, a proportional direction control valve, and an actuator (Fig. 10.1). Fluid power systems are widely used on aerospace, industrial, and mobile equipment because of their remarkable advantages over other control systems. The major advantages include high power-to-weight ratio, capability of being stalled, reversed, or operated intermittently, capability of fast response and acceleration, and reliable operation and long service life.

Due to differing tasks and working environments, the characteristics of fluid power systems are different for industrial and mobile applications (Lambeck, 1983). In industrial applications, low noise level is a major concern. Normally, a noise level below 70 dB is desirable and over 80 dB is excessive. Industrial systems commonly operate in the low (below 7 MPa or 1000 psi) to moderate (below 21 MPa or 3000 psi) pressure range. In mobile applications, the size is the premier concern. Therefore, mobile hydraulic systems commonly operate between 14 and 35 MPa (2000–5000 psi). Also, their allowable temperature operating range is usually higher than in industrial applications.

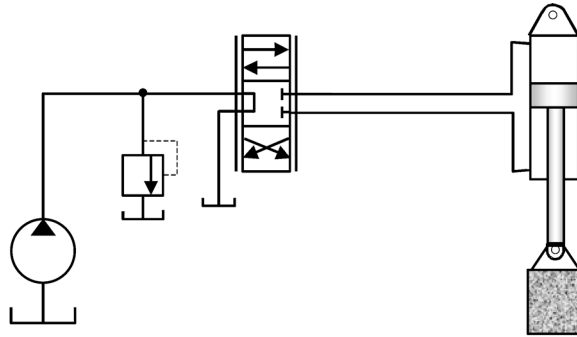



FIGURE 10.1 Schematic of a fluid power system.

Electrohydraulic Control Systems

The application of electronic controls to fluid power systems resulted in electrohydraulic control systems. Electrohydraulics has been widely used in aerospace, industrial, and mobile fluid power systems. Electrohydraulic controls have a few distinguishable advantages over other types of controls. First, an electrohydraulic system can be operated over a wide speed range, and its speed can be controlled continuously. More importantly, an electrohydraulic system can be stalled or operated under very large acceleration without causing its components to be damaged. A hydraulic actuator can be used in strong magnetic field without having the electromagnetic effects degrade control performance. In addition, hydraulic fluid flow can transfer heat away from system components and lubricate all moving parts continuously.

10.2 Hydraulic Fluids

Many types of fluids, e.g., mineral oils, biodegradable oils, and water-based fluids, are used in fluid power systems, depending on the task and the working environment. Ideally, hydraulic fluids should be inexpensive, noncorrosive, nontoxic, noninflammable, have good lubricity, and be stable in properties. The technically important properties of hydraulic fluids include density, viscosity, and bulk modulus.

Density

The density, ρ, of a fluid is defined as its mass per unit volume (Welty et al., 1984).

$$\rho = \frac{m}{V} \tag{10.1}$$

Density is approximately a linear function of pressure (P) and temperature (T) (Anderson, 1988).

$$\rho = \rho_0(1 + aP - bT) \tag{10.2}$$

In engineering practice, the manufacturers of the hydraulic fluids often provide the relative density (i.e., the specific gravity) instead of the actual density. The specific gravity of a fluid is the ratio of its actual density to the density of water at the same temperature.

Viscosity

The viscosity of a fluid is a measure of its resistance to deformation rate when subjected to a shearing force (Welty et al., 1984). Manufacturers often provide two kinds of viscosity values, namely the dynamic viscosity (μ) and the kinematic viscosity (ν). The dynamic viscosity is also named the absolute viscosity

and is defined by the Newtonian shear stress equation:

$$\mu = \frac{\tau}{\frac{dv}{dy}} \tag{10.3}$$

where dv is the relative velocity between two parallel layers dy apart, and τ is the shear stress.

The kinematic viscosity is the ratio of the dynamic viscosity to the density of the fluid and is defined using the following equation:

$$\nu = \frac{\mu}{\rho} \tag{10.4}$$

In the SI system, the unit of dynamic viscosity is Pascal-seconds (Pa s), and the unit of kinematic viscosity is square meter per second (m^2/s). Both the dynamic and kinematic vary strongly with temperature.

Bulk Modulus

Bulk modulus is a measure of the compressibility or the stiffness of a fluid. The basic definition of fluid bulk modulus is the fractional reduction in fluid volume corresponding to unit increase of applied pressure, expressed using the following equation (McCloy and Martin, 1973):

$$\beta = -V\left(\frac{\partial P}{\partial V}\right) \tag{10.5}$$

The bulk modulus can either be defined as the isothermal tangent bulk modulus if the compressibility is measured under a constant temperature or as the isentropic tangent bulk modulus if the compressibility is measured under constant entropy.

In analyzing the dynamic behavior of a hydraulic system, the stiffness of the hydraulic container plays a very important role. An effective bulk modulus, β_e, is often used to consider both the fluid's compressibility, β_f, and container stiffness, β_c, at the same time (Watton, 1989).

$$\frac{1}{\beta_e} = \frac{1}{\beta_f} + \frac{1}{\beta_c} \tag{10.6}$$

10.3 Hydraulic Control Valves

Principle of Valve Control

In a fluid power system, hydraulic control valves are used to control the pressure, flow rate, and flow direction. There are many ways to define a hydraulic valve so that a given valve can be named differently when it is used in different applications. Commonly, hydraulic valves can be classified based on their functions, such as pressure, flow, and directional control valves, or based on their control mechanisms, such as on-off, servo, and proportional electrohydraulic valves, or based on their structures, such as spool, poppet, and needle valves.

A hydraulic valve controls a fluid power system by opening and closing the flow-passing area of the valve. Such an adjustable flow-passing area is often described using an orifice area, A_o, in engineering practice. Physically, an orifice is a controllable hydraulic resistance, R_h. Under steady-state conditions, a hydraulic resistance can be defined as a ratio of pressure drop, Δp, across the valve to the flow rate, q, through the valve.

$$R_h = \frac{d(\Delta p)}{dq} \tag{10.7}$$

Control valves make use of many configurations of orifice to realize various hydraulic resistance characteristics for different applications. Therefore, it is essential to determine the relationship between the

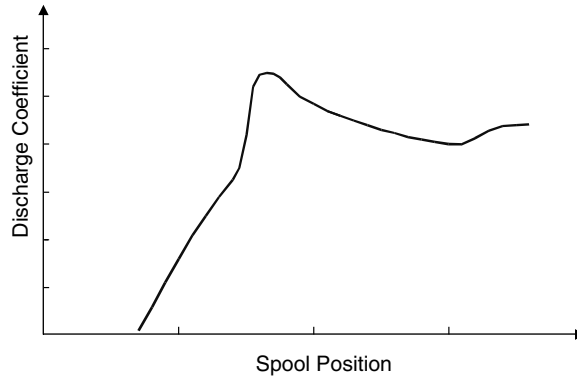



FIGURE 10.2 Discharge coefficient versus spool position in a spool valve.

pressure drop and the flow rate across the orifice. An orifice equation (McCloy and Martin, 1973) is often used to describe this relationship.

$$q = C_d A_o \sqrt{\frac{2}{\rho}\Delta P} \tag{10.8}$$

The pressure drop across the orifice is a system pressure loss in a fluid power system. In this equation, the orifice coefficient, C_d, plays an important role, and is normally determined experimentally. It has been found that the orifice coefficient varies greatly with the spool position, but does not appear to vary much with respect to the pressure drop across the orifice in a spool valve (Fig. 10.2, Viall and Zhang, 2000). Based on analytical results obtained from computational fluid dynamics simulations, the valve spool and sleeve geometries have little effect on the orifice coefficient for large spool displacements (Borghi et al., 1998).

Hydraulic Control Valves

There are many ways to classify hydraulic control valves. For instance, based on their structural configurations, hydraulic control valves can be grouped as cartridge valves and spool valves. This section will provide mathematical models of hydraulic control valves based on their structural configurations.

A typical cartridge valve has either a poppet or a ball to control the passing flow rate. Representing the control characteristics of a cartridge valve without loss of generality, a poppet type cartridge is analyzed (Fig. 10.3).

The control characteristics of a poppet type cartridge valve can be described using an orifice equation and a force balance equation. As shown in Fig. 10.3, the valve opens by lifting the poppet. Because of the cone structure of the poppet, the flow-passing area can be determined using the following equation:

$$A_x = \pi d x \sin\alpha \tag{10.9}$$

Therefore, the passing flow can be calculated using the orifice equation. For a poppet type valve, it is recommended to use a relative higher orifice coefficient of $c_d = 0.77 \sim 0.82$ (Li et al., 2000).

$$q = c_d A_x \sqrt{\frac{2}{\rho}(P_B - P_A)} \tag{10.10}$$

The forces acting on the poppet include the pressure, spring, and hydraulic forces. The pressure force can be determined based on the upstream, downstream, and spring chamber pressures.

$$F_P = P_A \frac{\pi d^2}{4} + P_B \frac{\pi(D^2 - d^2)}{4} - P_C \frac{\pi D^2}{4} \tag{10.11}$$

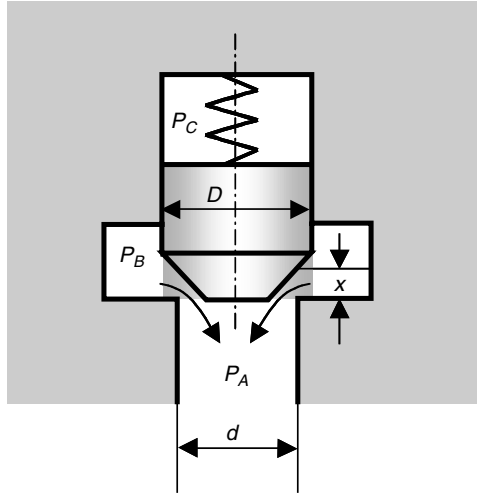



FIGURE 10.3 Operation principle of a puppet type cartridge valve.

The spring force biases the poppet towards closing. When the poppet is in the closed position, the spring force reaches its minimum value. The force increases as the poppet lifts to open the flow passage.

$$F_S = k(x_0 + x) \tag{10.12}$$

The steady-state flow force tends to open the poppet in this valve. The flow force is a function of the flow rate and fluid velocity passing through the valve orifice.

$$F_F = \rho q v \cos\alpha \tag{10.13}$$

The flow control characteristics of a spool valve are similar to those of a cartridge valve and can be described using an orifice equation. The only difference is that spool valve flow-passing area is determined by its wet perimeter, w, and spool displacement, x.

$$q = c_d w x \sqrt{\frac{2}{\rho}\Delta P} \tag{10.14}$$

If the orifice is formed by the edge of the spool and the valve body, the wet perimeter is $w = \pi d$. If the orifice is formed by n slots cut on the spool and the perimeter of each slot is n, the corresponding wet perimeter is $w = nb$. The orifice coefficient for a spool valve normally uses $c_d = 0.60\sim0.65$.

The forces acting on the spool also include the pressure, spring, and flow forces (Merritt, 1967). The pressure force is either balanced on the spool, because of its symmetric structure in a direct-actuator valve (actuated by a solenoid directly), or the pressure force to actuate the spool movement in a pilot actuated valve. The spring force tends to keep the spool in the central (neutral) position and can be described using Eq. (10.12). The flow forces acting on the spool can be calculated using Eq. (10.14). The flow velocity angle, α, is normally taken as 69°.

10.4 Hydraulic Pumps

Principles of Pump Operation

The pump is one of the most important components in a hydraulic system because it supplies hydraulic flow to the system. Driven by a prime mover, a hydraulic pump takes the fluid in at atmospheric pressure to fill an expanding volume of space inside the pump through an inlet port and delivers pressurized

fluids to the outlet due to the reduction in internal volume near the output port. The pump capacity is determined by pump displacement (D) and operating speed (n). The displacement of a pump is defined as the theoretical volume of fluid that can be delivered in one complete revolution of the pump shaft.

$$Q = Dn \tag{10.15}$$

The pump output pressure is determined by the system load, which is the combined resistance to fluid flow in the pipeline and the resistance to move an external load. Unless the pump flow has egress either by moving a load or by passing through a relief valve back to the reservoir, excessive pressure build-up can cause serious damage to the pump and/or the connecting pipeline (Reed and Larman, 1985).

Based on their ability to change displacement, hydraulic pumps can be categorized as fixed-flow or variable-flow pumps. Based on their design, hydraulic pumps can be categorized as gear pumps, vane pumps, and piston pumps. Normally, gear pumps are fixed-flow pumps, and vane pumps and piston pumps can be either fixed-flow pumps or variable-flow pumps.

The choice of pump design varies from industry to industry. For example, the machine tool manufacturers often select vane pumps because of their low noise, and their capability to deliver a variable flow at a constant pressure. Mobile equipment manufacturers like to use piston pumps due to their high power-to-weight ratio. Some agricultural equipment manufacturers prefer gear pumps for their low cost and robustness (Reed and Larman, 1985), but piston pumps are also popular.

Pump Controls and Systems

Pumps are energy conversion devices that convert mechanical energy into fluid potential energy to drive various hydraulic actuators to do work. To meet the requirements of different applications, there are many types of fluid power system controls from which to choose. The design of the directional control valve must be compatible with the pump design. Normally, an open-center directional control valve is used with a fixed displacement pump and a closed-center directional control valve is used in a circuit equipped with a variable displacement pump.

A fluid power system including a fixed displacement pump and an open-center directional control valve (Fig. 10.1) is an open-loop open-center system. Such a system is also called a load-sensitive system because the pump delivers only the pressure required to move the load, plus the pressure drop to overcome line losses. The open-loop open-center system is suitable for simple "on-off" controls. In such operations, the hydraulic actuator either moves the load at the maximum velocity or remains stationary with the pump unloaded. If a proportional valve is used, the open-loop open-center system can also achieve velocity control of the actuator. However, such control will increase the pressure of the extra flow for releasing it back to the tank. Such control causes significant power loss and results in low system efficiency and heat generation.

To solve this problem, an open-loop closed-center circuit is constructed using a variable displacement pump and a closed-center directional control valve. Because a variable displacement pump is commonly equipped with a pressure-limiting control or "pressure compensator," the pump displacement will be automatically increased or decreased as the system pressure decreases or increases. If the metering position of the directional control valve is used to control the actuator velocity, constant velocity can be achieved if the load is constant. However, if the load is changing, the "pressure-compensating" system will not be able to keep a constant velocity without adjusting the metering position of the control valve. To solve this problem, a "load-sensing" pump should be selected for keeping a constant velocity under changing load. The reason for a "load-sensing" pump being able to maintain a constant velocity for any valve-metering position is that it maintains a constant pressure drop across the metering orifice of the directional control valve, and automatically adjusts the pump outlet pressure to compensate for the changes in pressure caused by external load. The constant pressure drop across the valve maintains constant flow, and therefore, constant load velocity.

10.5 Hydraulic Cylinders

A hydraulic cylinder transfers the potential energy of the pressurized fluid into mechanical energy to drive the operating device performing linear motions and is the most common actuator used in hydraulic systems. A hydraulic cylinder consists of a cylinder body, a piston, a rod, and seals. Based on their structure, hydraulic cylinders can be classified as single acting (applying force in one direction only), double acting (exerts force in either direction), single rod (does not have a rod at the cap side), and double rod (has a rod at both sides of the piston) cylinders.

Cylinder Parameters

A hydraulic cylinder transfers energy by converting the flow rate and pressure into the force and velocity. The velocity and the force from a double-acting double-rod cylinder can be determined using the following equations:

$$v = \frac{4q}{\pi(D^2 - d^2)} \tag{10.16}$$

$$F = \frac{\pi}{4}(D^2 - d^2)(P_1 - P_2) \tag{10.17}$$

The velocity and the force from a double-acting single-rod cylinder should be determined differently for extending and retracting motions. In retraction, the velocity can be determined using Eq. (10.16), and the force can be determined using the following equation:

$$F = P_1\frac{\pi(D^2 - d^2)}{4} - P_2\frac{\pi D^2}{4} \tag{10.18}$$

In extension, the velocity and exerting forces can be determined using the following equations:

$$v = \frac{4q}{\pi D^2} \tag{10.19}$$

$$F = (P_1 - P_2)\frac{\pi D^2}{4} + P_2\frac{\pi d^2}{4} \tag{10.20}$$

The hydraulic stiffness, k_h, of the cylinder plays an important role in the dynamic performance of a hydraulic system. It is a function of fluid bulk modulus (β), piston areas (A_1, A_2), cylinder chamber volumes (V_1, V_2), and the volume of hydraulic hoses connected to both chambers (V_{L1}, V_{L2}). For a double-acting single-rod cylinder, the stiffness on both sides of the piston acts in parallel (Skinner and Long, 1998). The total stiffness of the cylinder is given by the following equation:

$$k_h = \beta\left(\frac{A_1^2}{V_{L1} + V_1} + \frac{A_1^2}{V_{L2} + V_2}\right) \tag{10.21}$$

The natural frequency, ω_n, of a hydraulic system is determined by the combined mass, m, of the cylinder and the load using the following equation:

$$\omega_n = \sqrt{\frac{k_h}{m}} \tag{10.22}$$

10.6 Fluid Power Systems Control

System Steady-State Characteristics

The steady-state characteristics of a fluid power system determine loading performance, speed control capability, and the efficiency of the system. Modeling a hydraulic system without loss of generality, a system consisting of an open-center four-way directional control valve and a single-rod double acting cylinder is used to analyze the steady-state characteristics of the system (Fig. 10.1). In this system, the orifice area of the cylinder-to-tank (C-T) port in the control valve is always larger than that of the pump-to-cylinder (P-C) port. Therefore, it is reasonable to assume that the P-C orifice controls the cylinder speed during extension (Zhang, 2000).

Based on Newton's Law, the force balance on the piston is determined by the head-end chamber pressure, P_1, the head-end piston area, A_1, the rod-end chamber pressure, P_2, the rod-end piston area, A_2, and the external load, F, when the friction and leakage are neglected.

$$P_1A_1 - P_2A_2 = F \tag{10.23}$$

If neglecting the line losses from actuator to reservoir, the rod-end pressure equals zero. Then, the head-end pressure is determined by the external load to the system.

$$P_1 = \frac{F}{A_1} \tag{10.24}$$

In order to push the fluid passing the control valve and entering the head-end of the cylinder, the discharge pressure, P_P, of the hydraulic pump has to be higher than the cylinder chamber pressure. The difference between the pump discharge pressure and the cylinder chamber pressure is determined by the hydraulic resistance across the control valve. Based on the orifice equation, the flow rate entering the cylinder head-end chamber is

$$q = C_dA_o\sqrt{\frac{2}{\rho}(P_P - P_1)} \tag{10.25}$$

Using a control coefficient, K, to represent C_d and ρ, the cylinder speed can be described using the following equation:

$$v = \frac{KA_o}{A_1}\sqrt{P_P - \frac{F}{A_1}} \tag{10.26}$$

Equation (10.13) describes the speed-load relationship of a hydraulic cylinder under a certain fluid passing area (orifice area) of the control valve. Depicted in Fig. 10.4, the cylinder speed decreases as the external load applied to the cylinder increases. When there is no external load, the cylinder speed reaches a maximum. Conversely, when the external load researches the valve of $F = P_PA_1$, then the cylinder will stall. The stall load is independent of the size of the fluid passing area in the valve. Such characteristics of a fluid power system eliminate the potential of overloading, which makes it a safer power transmission method.

In system analysis, the speed stiffness, k_v, is often used to describe the consistency of the cylinder speed under changing system load (Li et al., 2000).

$$k_v = -\frac{1}{\frac{\partial v}{\partial F}} = \frac{2(P_PA_1 - F)}{v} \tag{10.27}$$

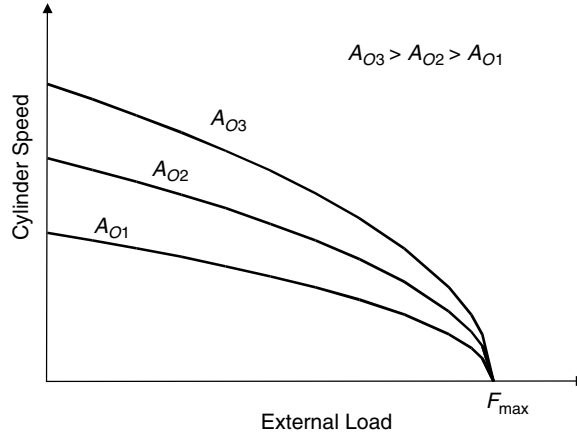



FIGURE 10.4 Hydraulic cylinder load-speed relationship under the same system pressure.

Equation (10.27) indicates that the increase in speed stiffness can be achieved either by increasing the system pressure or the cylinder size, or by decreasing the speed.

System Dynamic Characteristics

To analyze the dynamic characteristics of this hydraulic cylinder actuation system, one can use flow continuity and system momentum equations to model the cylinder motion. Neglecting system leakage, friction, and line loss, the following are the governing equations for the hydraulic system:

$$q = kx\sqrt{P_P - P_1} = A_1\frac{dy}{dt} + \frac{V_1}{\beta}\frac{dP_1}{dt} \tag{10.28}$$

$$P_1A_1 = m\frac{d^2y}{dt^2} + F \tag{10.29}$$

To perform dynamic analysis on this hydraulic system, it is essential to derive its transfer function based on the above nonlinear equation, which can be obtained by taking the Laplace transform on the linearized form of the above equations (Watton, 1989).

$$\delta v(s) = \frac{\frac{k_1K_i}{A_1}\delta i(s) - \left(\frac{V_1}{A_1^2\beta}s + \frac{1}{A_1^2k_3R_o}\right)\delta F(s)}{\frac{V_1}{A_1^2\beta}ms^2 + \frac{1}{A_1^2k_2R_o}ms + 1} \tag{10.30}$$

Making

$$\omega_n = \sqrt{\frac{A_1^2\beta}{V_1m}}, \quad \zeta = \frac{1}{2k_2R_o}\sqrt{\frac{m\beta}{V_1A_1^2}}, \quad \text{and} \quad K_s = \frac{k_1K_i}{A_1}$$

Equation (10.30) can be represented as

$$\delta v(s) = \frac{K_s\delta i(s)}{\frac{1}{\omega_n^2}s^2 + \frac{2\zeta}{\omega_n}s + 1} - \frac{\frac{1}{A_1^2}\left(\frac{V_1}{\beta}s + \frac{1}{k_3R_o}\right)\delta F(s)}{\frac{1}{\omega_n^2}s^2 + \frac{2\zeta}{\omega_n}s + 1} \tag{10.31}$$

Based on the stability criterion for a second-order system, it should satisfy

$$\frac{1}{W_n^2}s^2 + \frac{2Z}{W_n}s + 1 = 0 \quad (10.32)$$

The speed control coefficient, K_s, is the gain between the control signal current and the cylinder speed. A higher gain can increase the system sensitivity in speed control.

E/H System Feedforward-Plus-PID Control

Equation (10.31) indicates that the speed control of a hydraulic cylinder is a third-order system. Its dynamic behaviors are affected by spool valve characteristics, system pressure, and cylinder size. Therefore, it is a challenging job to realize accurate and smooth speed control on a hydraulic cylinder. A feedforward plus proportional integral derivative (FPID) controller has proven capable of achieving high-speed control performance of a hydraulic cylinder (Zhang, 1999).

An FPID controller consists of a feedforward loop and a PID loop (Fig. 10.5). The feedforward loop is designed to compensate for the nonlinearity of the hydraulic system, including the deadband of the system and the nonlinear flow gain of the control valve. It uses a feedforward gain to determine the basic control input based on demand speed. This feedforward gain is scheduled based on the inverse valve transform, which provides the steady-state control characteristics of the E/H valve in terms of cylinder speed and control-current to valve PWM driver.

The PID loop complements the feedforward control via the speed tracking error compensation. The PID controller is developed based on the transfer function of the linearized system for the hydraulic cylinder speed control system.

$$G(s) = \left(K_P + \frac{K_I}{s} + K_D s\right) \quad (10.33)$$

The robustness of the FPID control was evaluated based on its performance and stability. Performance robustness deals with unexpected external disturbances and stability robustness deals with internal structural or parametric changes in the system. The design of this FPID controller was based on a worst-case scenario of system operating conditions in tuning both the PID gains and the feedforward gain.

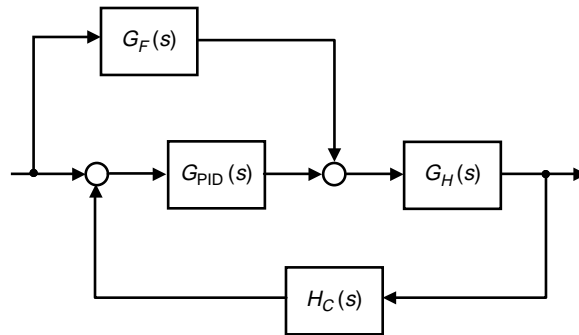



FIGURE 10.5 Schematic block diagram of the feedforward-plus-PID controller. $G_F(s)$ is the feedforward gain, $G_{PID}(s)$ is the overall gain of the feedback PID controller, $G_H(s)$ is hydraulic system gain, and $H_C(s)$ is the sensor gain.

E/H System Generic Fuzzy Control

Fuzzy control is an advanced control technology that can mimic a human's operating strategy in controlling complex systems and can handle systems with uncertainty and nonlinearity (Pedrycz, 1993). One common feature of fuzzy controllers is that most such controllers are designed based on natural language control laws. This feature makes it possible to design a generic controller for different plants if the control of those plants can be described using the same natural language control laws (Zhang, 2001).

The speed control on a hydraulic cylinder actually is achieved by regulating the supplied flow rate to the cylinder. In different hydraulic systems, the size of the cylinder and the capability of hydraulic system are usually different, but the control principles are very similar. Representing cylinder speed control operation, using natural language without loss of generality, the control laws are the same for all systems:

To have a fast motion, open the valve fully.
To make a slow motion, keep the valve open a_little.
To hold the cylinder at its current position, return the valve to the center.
To make a reverse motion, operate the valve to the other direction.

This natural language model represents the general roles in controlling the cylinder speed via an E/H control valve on all hydraulic systems. The differences in system parameters on different systems can be handled by redefining the domain of the fuzzy variable, such as fully, a_lot, and a_little, using fuzzy membership functions (Passino and Yurkovich, 1998). This model provides the basis for designing a generic fuzzy controller for E/H systems. The adoption of the generic controller on different systems can be as easy as redefining the fuzzy membership function based on its system parameters.

Figure 10.6 shows the block diagram of a generic fuzzy controller consisting of two input variable fuzzifiers, a control rule base, and a control command defuzzifier. The two input fuzzifiers were designed to convert real-valued input variables into linguistic variables with appropriate fuzzy memberships. Each fuzzifier consists of a set of fuzzy membership functions defining the domain for each linguistic input variable. A real-valued input variable is normally converted into two linguistic values with associated memberships. The definitions of these fuzzy values play a critical role in the design of generic fuzzy controllers and are commonly defined based upon hydraulic system parameters. The fuzzy controller uses fuzzy control rules to determine control actions according to typical behaviors in the speed control of hydraulic cylinders. The control outputs are also linguistic values and associated with fuzzy memberships. For example, if the demanding speed is negative_small (NS) and the error in speed was positive_small (PS), the appropriate valve control action will be positive_small (PS).

The appropriate control actions were determined based on predefined control rules. Since each real-valued variable commonly maps into two fuzzy values, the fuzzy inference engine fires at least two control rules containing these fuzzy values to determine the appropriate control action. Therefore, at least two appropriate fuzzy-valued control actions will be selected. However, the E/H controller can only implement one specific real-value control command at a given time. It is necessary to convert multiple fuzzy-valued control commands into one real-valued control signal in this fuzzy controller.

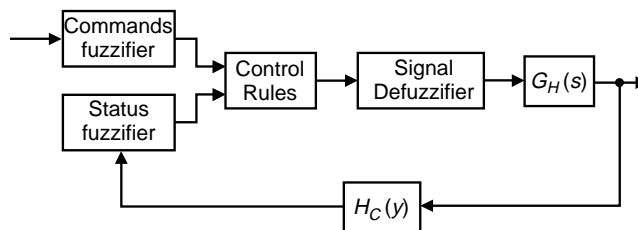



FIGURE 10.6 Block diagram of fuzzy E/H control system. The fuzzy controller consists of input variable fuzzifiers, control rules, and a signal defuzzifier.

The defuzzification process converts two or more fuzzy-valued outputs to one real-valued output. There are many defuzzification methods, such as center of gravity (COG) and center of area (COA), available for different applications (Passino and Yurkovich, 1998). By COA approach, the real-valued control signal, u, was determined by the domain and the memberships of the selected fuzzy control commands, $\mu(u_i)$, using the following equation:

$$u = \frac{\sum_{i=1}^{n} u_i \mu(u_i) du}{\sum_{i=1}^{n} \mu(u_i) du} \tag{10.34}$$

The COA method naturally averages the domains of selected fuzzy control commands, and thus reduces the sensitivity of the system to noise. The use of a COA approach increased the robustness and accuracy of the control.

The performance of the fuzzy controller depends on the appropriation of domain definition for both input and output fuzzy variables. Properly defined fuzzy variables for a specific E/H system will improve the stability, accuracy, and nonlinearity compensation of the fuzzy controller. Normally, a triangular fuzzy membership function, μ_{FV}, was defined by domain values of a_i, a_j, and a_k, for each fuzzy value (FV) in the fuzzy controller.

$$\mu_A = \begin{bmatrix} \mu_{NL} \\ \mu_{NM} \\ \mu_{NS} \\ \mu_{ZE} \\ \mu_{PS} \\ \mu_{PM} \\ \mu_{PL} \end{bmatrix} = \begin{bmatrix} a_1 & a_1 & a_2 \\ a_1 & a_2 & a_3 \\ a_2 & a_3 & a_4 \\ a_3 & a_4 & a_5 \\ a_4 & a_5 & a_6 \\ a_5 & a_6 & a_7 \\ a_6 & a_7 & a_7 \end{bmatrix} \tag{10.35}$$

where μ_A is a set of fuzzy membership functions for each fuzzy input or output variable; a_t, a_k are the boundaries; and a_j is the full membership point of the fuzzy value.

Equation (10.35) uses a set of seven domain values to define seven fuzzy values in the real-valued operating range. The tuning of the fuzzy controller was to determine the domain values for each of the fuzzy values. The following vector presents the domains of fuzzy membership functions for a particular variable:

$$A = \{a_1 \;\; a_2 \;\; a_3 \;\; a_4 \;\; a_5 \;\; a_6 \;\; a_7\} \tag{10.36}$$

10.7 Programmable Electrohydraulic Valves

Proportional directional control valves are by far the most common means for motion control of hydraulic motors or cylinders in fluid power systems (McCloy, 1973). Normally, a proportional direction control valve uses a sliding spool to control the direction and the amount of fluid passing through the valve. For different applications, the spool in a proportional direction control valve is often specially designed to provide the desired control characteristics. As a result, valves are specific and cannot be interchangeable even if they are exactly of the same size. The multiplicity of such specific valves make them inconvenient and costly to manufacture, distribute, and service. To provide a solution to these problems, researchers at the University of Illinois at Urbana-Champaign (Book and Goering, 1999; Hu et al., 2001) developed a generic programmable electrohydraulic (E/H) control valve. A generic programmable valve is a set of individually

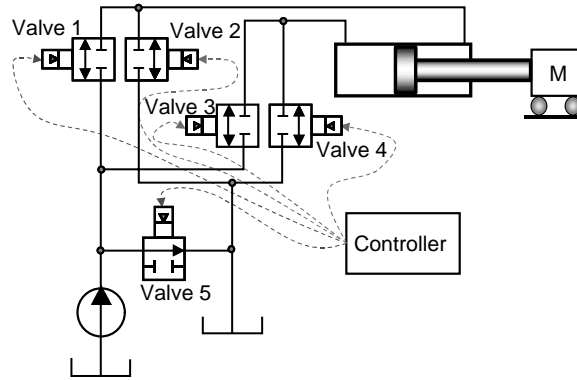



FIGURE 10.7 System schematic of a hydraulic system using generic programmable E/H valves.

controlled E/H valves capable of fulfilling flow and pressure control requirements. One set of such generic valves can replace a proportional direction control valve and other auxiliary valves, such as line release valves, in a circuit.

A generic programmable E/H valve is normally constructed using five bi-directional, proportional flow control sub-valves, three pressure sensors, and an electronic controller. Figure 10.7 shows the schematic of the generic valve circuit. Sub-valves 1 and 2 connect the pump and the head-end or the rod-end chambers of the cylinder and provide equilibrium ports of P-to-A and P-to-B as in a conventional direction control valve, while sub-valves 3 and 4 connect cylinder chambers A or B to the tank and provide equilibrium ports of A-to-T and B-to-T of a direction control valve. Sub-valve 5 connects the pump and the tank directly and provides a dual-function of line release and an equilibrium port of P-to-T of a direction control valve. By controlling the opening and closing of these sub-valves, the basic functions of the generic valve can be realized. In operation, the controller output control signals for each sub-valve are based on a predefined control logic.

With proper logic in the on-off control of all five sub-valves, the generic programmable valve was capable of realizing several basic functions, including open-center, closed-center, float-center, make-up, and pressure release functions. By applying modulation control, the generic valve can realize proportional functions such as meter-in/meter-out, load sensing, regeneration, and anti-cavitation. For example, in a conventional tandem-center or closed-center direction control valve, the ports A and B are normally closed for holding the pressure in cylinder chambers, while the ports P and T are either normally open or closed. To fulfill this function, the generic valve keeps sub-valves 1–4 closed to hold the cylinder chamber pressure, and fully opens sub-valve 5 to bleed the flow back to the tank, either at low pressure (tandem-center function) or when the system pressure exceeds a preset level (closed-center function). In conventional open-center direction control valves, all ports are normally connected. To fulfill this function, the generic valve keeps all sub-valves open. Similarly, to provide float-center function, the generic valve needs to open sub-valves 3 and 4 to release pressure in both the head-end and the rod-end chambers of the cylinder. In both cases, sub-valve 5 will be opened only when the system pressure exceeds a preset level.

It is almost impossible to achieve the regeneration function from a conventional direction control valve. In achieving this function, a generic valve needs to open sub-valves 1 and 2 to lead the returning flow of the rod-end chamber back to the head-end chamber to provide additional flow for increasing the extending speed. Make-up function in a conventional hydraulic system is provided by a separate make-up valve for supplying fluid directly from the tank in case of cavitation. The generic valve can also provide this function by actuating the corresponding cylinder-to-tank sub-valves open when the system pressure is below a certain level.

References

1. Anderson, W.R., *Controlling Electrohydraulic Systems,* Marcel Dekker, New York, NY, 1988.
2. Book, R. and Goering, C.E., Programmable electrohydraulic valve, *SAE 1999 Transactions, Journal of Commercial Vehicles* (1997), Section 2, 108:346–352.
3. Borghi, M.G., Cantore, G., Milani, M., and Paoluzzi, R., Analysis of hydraulic components using computational fluid dynamics models, *Proceedings of the Institution of Mechanical Engineers, Journal C* (1998), 212:619–629.
4. Lambeck, R.P., *Hydraulic Pumps and Motors: Selection and Application for Hydraulic Power Control Systems,* Marcel Dekker, New York, NY, 1983.
5. Li, Z., Ge, Y., and Chen, Y., *Hydraulic Components and Systems* (in Chinese), Mechanical Industry Publishing, Beijing, China, 2000.
6. Hu, H., Zhang, Q., and Alleyne, A., Multi-function realization of a generic programmable E/H valve using flexible control logic, *Proceedings of the Fifth International Conference on Fluid Power Transmission and Control* (2001), International Academic Publishers, Beijing, China, pp. 107–110.
7. Merritt, H.E., *Hydraulic Control Systems,* John Wiley & Sons, New York, NY, 1967.
8. McCloy, D. and Martin, H.R., *The Control of Fluid Power,* John Wiley & Sons, New York, NY, 1973.
9. Passino, K.M. and Yurkovich, S., *Fuzzy Control*, Addition-Wesley, Menlo Park, CA, 1998.
10. Pedrycz, W., *Fuzzy Control and Fuzzy Systems,* 2nd edition, Wiley, New York, NY, 1993.
11. Reed, E.W. and Larman, I.S., *Fluid Power with Microprocessor Control: An Introduction,* Prentice-Hall, New York, NY, 1985.
12. Skinner, S.C. and Long, R.J., *Closed Loop Electrohydraulic Systems Manual,* 2nd edition, Vickers, Rochester Hills, MI, 1998.
13. Viall, E.N. and Zhang, Q., Determining the discharge coefficient of a spool valve, *Proceedings of the American Control Conference* (2000), Chicago, IL, pp. 3600–3604.
14. Watton, J., *Fluid Power Systems, Modeling, Simulation, Analog and Microcomputer Control,* Prentice-Hall, New York, NY, 1989.
15. Welty, J.R., Wicks, C.E., and Wilson, R.E., *Fundamentals of Momentum, Heat, and Mass Transfer,* 3rd edition, John Wiley & Sons, New York, NY, 1984.
16. Zhang, Q., Hydraulic linear actuator velocity control using a feedforward-plus-PID control, *International Journal of Flexible Automation and Integrated Manufacturing* (1999), 7:275–290.
17. Zhang, Q., Design of a generic fuzzy controller for electrohydraulic steering, *Proceedings of the American Control Conference* (2001), (in press).

11
Electrical Engineering

Giorgio Rizzoni
Ohio State University

11.1 Introduction

The role played by electrical and electronic engineering in mechanical systems has dramatically increased in importance in the past two decades, thanks to advances in integrated circuit electronics and in materials that have permitted the integration of sensing, computing, and actuation technology into industrial systems and consumer products. Examples of this integration revolution, which has been referred to as a new field called *Mechatronics*, can be found in consumer electronics (auto-focus cameras, printers, microprocessor-controlled appliances), in industrial automation, and in transportation systems, most notably in passenger vehicles. The aim of this chapter is to review and summarize the foundations of electrical engineering for the purpose of providing the practicing mechanical engineer a quick and useful reference to the different fields of electrical engineering. Special emphasis has been placed on those topics that are likely to be relevant to product design.

11.2 Fundamentals of Electric Circuits

This section presents the fundamental laws of circuit analysis and serves as the foundation for the study of electrical circuits. The fundamental concepts developed in these first pages will be called on through the chapter.

The fundamental electric quantity is **charge**, and the smallest amount of charge that exists is the charge carried by an electron, equal to

$$q_e = -1.602 \times 10^{-19} \text{ coulomb} \tag{11.1}$$

As you can see, the amount of charge associated with an electron is rather small. This, of course, has to do with the size of the unit we use to measure charge, the **coulomb** (C), named after Charles Coulomb. However, the definition of the coulomb leads to an appropriate unit when we define electric current,

since current consists of the flow of very large numbers of charge particles. The other charge-carrying particle in an atom, the proton, is assigned a positive sign and the same magnitude. The charge of a proton is

$$q_p = +1.602 \times 10^{-19} \text{ coulomb} \tag{11.2}$$

Electrons and protons are often referred to as **elementary charges**.

Electric current is defined as the time rate of change of charge passing through a predetermined area. If we consider the effect of the enormous number of elementary charges actually flowing, we can write this relationship in differential form:

$$i = \frac{dq}{dt} \text{ (C/sec)} \tag{11.3}$$

The units of current are called **amperes** (A), where 1 A = 1 C/sec. The electrical engineering convention states that the positive direction of current flow is that of positive charges. In metallic conductors, however, current is carried by negative charges; these charges are the free electrons in the conduction band, which are only weakly attracted to the atomic structure in metallic elements and are therefore easily displaced in the presence of electric fields.

In order for current to flow there must exist a closed circuit. Figure 11.1 depicts a simple circuit, composed of a battery (e.g., a dry-cell or alkaline 1.5-V battery) and a light bulb.

Note that in the circuit of Fig. 11.1, the current, i, flowing from the battery to the resistor is equal to the current flowing from the light bulb to the battery. In other words, no current (and therefore no charge) is “lost” around the closed circuit. This principle was observed by the German scientist G.R. Kirchhoff and is now known as **Kirchhoff’s current law** (KCL). KCL states that because charge cannot be created but must be conserved, *the sum of the currents at a node must equal zero* (in an electrical circuit, a **node** is the junction of two or more conductors). Formally:

$$\sum_{n=1}^{N} i_n = 0 \quad \text{Kirchhoff's current law} \tag{11.4}$$

The significance of KCL is illustrated in Fig. 11.2, where the simple circuit of Fig. 11.2 has been augmented by the addition of two light bulbs (note how the two nodes that exist in this circuit have been emphasized by the shaded areas). In applying KCL, one usually defines currents entering a node as being negative and currents exiting the node as being positive. Thus, the resulting expression for the circuit of Fig. 11.2 is

$$i + i_1 + i_2 + i_3 = 0$$

Charge moving in an electric circuit gives rise to a current, as stated in the preceding section. Naturally, it must take some work, or energy, for the charge to move between two points in a circuit, say, from point a to point b. The total *work per unit charge* associated with the motion of charge between two

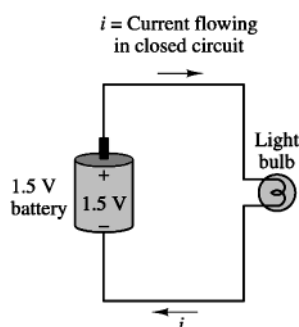



FIGURE 11.1 A simple electrical circuit.

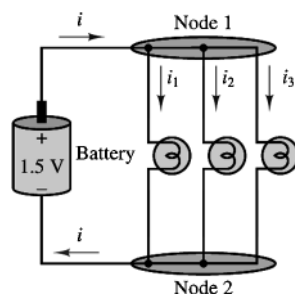



FIGURE 11.2 Illustration of Kirchhoff's current law.

FIGURE 11.3 Voltages around a circuit.

points is called **voltage**. Thus, the units of voltage are those of energy per unit charge:

$$1 \text{ volt} = \frac{1 \text{ joule}}{\text{coulomb}} \tag{11.5}$$

The voltage, or **potential difference**, between two points in a circuit indicates the energy required to move charge from one point to the other. As will be presently shown, the direction, or polarity, of the voltage is closely tied to whether energy is being dissipated or generated in the process. The seemingly abstract concept of work being done in moving charges can be directly applied to the analysis of electrical circuits; consider again the simple circuit consisting of a battery and a light bulb. The circuit is drawn again for convenience in Fig. 11.3, and nodes are defined by the letters a and b. A series of carefully conducted experimental observations regarding the nature of voltages in an electric circuit led Kirchhoff to the formulation of the second of his laws, **Kirchhoff's voltage law**, or KVL. The principle underlying KVL is that no energy is lost or created in an electric circuit; in circuit terms, the sum of all voltages associated with sources must equal the sum of the load voltages, so that *the net voltage around a closed circuit is zero.* If this were not the case, we would need to find a physical explanation for the excess (or missing) energy not accounted for in the voltages around a circuit. KVL may be stated in a form similar to that used for KCL:

$$\sum_{n=1}^{N} v_n = 0 \quad \text{Kirchhoff's voltage law} \tag{11.6}$$

where the v_n are the individual voltages around the closed circuit. Making reference to Fig. 11.3, we can see that it must follow from KVL that the work generated by the battery is equal to the energy dissipated in the light bulb to sustain the current flow and to convert the electric energy to heat and light:

$$v_{ab} = -v_{ba}$$

or

$$v_1 = v_2$$

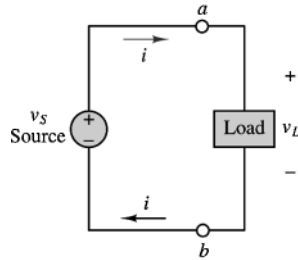



FIGURE 11.4 Sources and loads in an electrical circuit.

One may think of the work done in moving a charge from point a to point b and the work done moving it back from b to a as corresponding directly to the *voltages across individual circuit elements.* Let Q be the total charge that moves around the circuit per unit time, giving rise to the current i. Then the work done in moving Q from b to a (i.e., across the battery) is

$$W_{ba} = Q \times 1.5 \text{ V} \tag{11.7}$$

Similarly, work is done in moving Q from a to b, that is, across the light bulb. Note that the word *potential* is quite appropriate as a synonym of voltage, in that voltage represents the potential energy between two points in a circuit: if we remove the light bulb from its connections to the battery, there still exists a voltage across the (now disconnected) terminals b and a.

A moment's reflection upon the significance of voltage should suggest that it must be necessary to specify a sign for this quantity. Consider, again, the same dry-cell or alkaline battery, where, by virtue of an electrochemically induced separation of charge, a 1.5-V potential difference is generated. The potential generated by the battery may be used to move charge in a circuit. The rate at which charge is moved once a closed circuit is established (i.e., the current drawn by the circuit connected to the battery) depends now on the circuit element we choose to connect to the battery. Thus, while the voltage across the battery represents the potential for *providing energy* to a circuit, the voltage across the light bulb indicates the amount of work done in *dissipating energy.* In the first case, energy is generated; in the second, it is consumed (note that energy may also be stored, by suitable circuit elements yet to be introduced). This fundamental distinction required attention in defining the sign (or polarity) of voltages.

We shall, in general, refer to elements that provide energy as **sources**, and to elements that dissipate energy as **loads**. Standard symbols for a generalized source-and-load circuit are shown in Fig. 11.4. Formal definitions will be given in a later section.

Electric Power and Sign Convention

The definition of voltage as work per unit charge lends itself very conveniently to the introduction of power. Recall that power is defined as the work done per unit time. Thus, the power, P, either generated or dissipated by a circuit element can be represented by the following relationship:

$$\text{Power} = \frac{\text{work}}{\text{time}} = \frac{\text{work}}{\text{unit charge}}\frac{\text{charge}}{\text{time}} = \text{voltage} \times \text{current} \tag{11.8}$$

Thus, the electrical power generated by an active element, or that dissipated or stored by a passive element, is equal to the product of the voltage across the element and the current flowing through it.

$$P = VI \tag{11.9}$$

It is easy to verify that the units of voltage (joules/coulomb) times current (coulombs/second) are indeed those of power (joules/second, or watts).

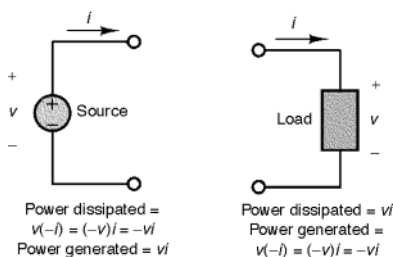



FIGURE 11.5 The passive sign convention.

It is important to realize that, just like voltage, power is a signed quantity, and that it is necessary to make a distinction between *positive* and *negative power.* This distinction can be understood with reference to Fig. 11.5, in which a source and a load are shown side by side. The polarity of the voltage across the source and the direction of the current through it indicate that the voltage source *is doing work in moving charge from a lower potential to a higher potential.* On the other hand, the load is dissipating energy, because the direction of the current indicates that *charge is being displaced from a higher potential to a lower potential.* To avoid confusion with regard to the sign of power, the electrical engineering community uniformly adopts the **passive sign convention**, which simply states that *the power dissipated by a load is a positive quantity* (or, conversely, that the power generated by a source is a positive quantity). Another way of phrasing the same concept is to state that if current flows from a higher to a lower voltage (+ to –), the power dissipated will be a positive quantity.

Circuit Elements and Their *i-v* Characteristics

The relationship between current and voltage at the terminals of a circuit element defines the behavior of that element within the circuit. In this section, we shall introduce a graphical means of representing the terminal characteristics of circuit elements. Figure 11.6 depicts the representation that will be employed throughout the chapter to denote a generalized circuit element: the variable i represents the current flowing through the element, while v is the potential difference, or voltage, across the element.

FIGURE 11.6 Generalized representation of circuit elements.

Suppose now that a known voltage were imposed across a circuit element. The current that would flow as a consequence of this voltage, and the voltage itself, form a unique pair of values. If the voltage applied to the element were varied and the resulting current measured, it would be possible to construct a functional relationship between voltage and current known as the ***i-v* characteristic** (or **volt-ampere characteristic**). Such a relationship defines the circuit element, in the sense that if we impose any prescribed voltage (or current), the resulting current (or voltage) is directly obtainable from the *i-v* characteristic. A direct consequence is that the power dissipated (or generated) by the element may also be determined from the *i-v* curve.

The *i-v* characteristics of ideal current and voltage sources can also be useful in visually representing their behavior. An ideal voltage source generates a prescribed voltage independent of the current drawn from the load; thus, its *i-v* characteristic is a straight vertical line with a voltage axis intercept corresponding to the source voltage. Similarly, the *i-v* characteristic of an ideal current source is a horizontal line with a current axis intercept corresponding to the source current. Figure 11.7 depicts this behavior.

Resistance and Ohm's Law

When electric current flows through a metal wire or through other circuit elements, it encounters a certain amount of **resistance,** the magnitude of which depends on the electrical properties of the material. Resistance to the flow of current may be undesired—for example, in the case of lead wires and connection

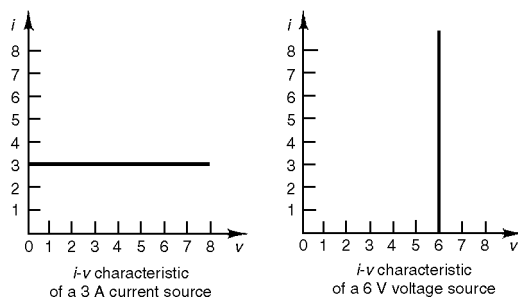



FIGURE 11.7 *i-v* characteristics of ideal sources.

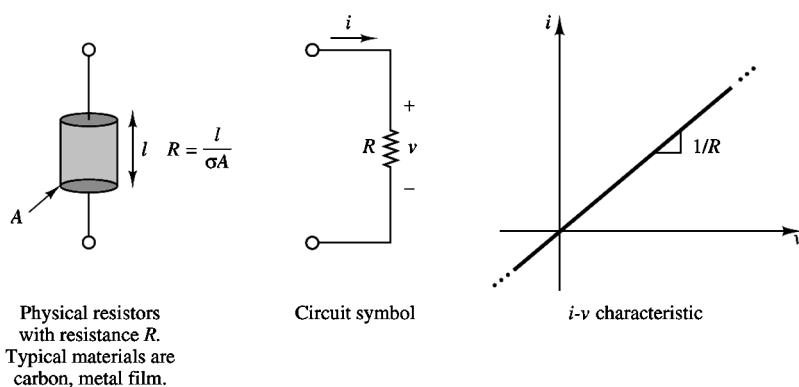



FIGURE 11.8 The resistance element.

cable—or it may be exploited in an electrical circuit in a useful way. Nevertheless, practically all circuit elements exhibit some resistance; as a consequence, current flowing through an element will cause energy to be dissipated in the form of heat. An ideal **resistor** is a device that exhibits linear resistance properties according to Ohm's law, which states that

$$V = IR \tag{11.10}$$

that is, that the voltage across an element is directly proportional to the current flow through it. R is the value of the resistance in units of ohms (Ω), where

$$1\ \Omega = 1\ \text{V/A} \tag{11.11}$$

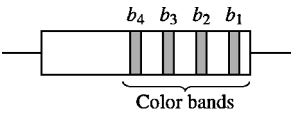
The resistance of a material depends on a property called **resistivity**, denoted by the symbol ρ; the inverse of resistivity is called **conductivity** and is denoted by the symbol σ. For a cylindrical resistance element (shown in Fig. 11.8), the resistance is proportional to the length of the sample, l, and inversely proportional to its cross-sectional area, A, and conductivity, σ.

$$v = \frac{1}{\sigma A} i \tag{11.12}$$

It is often convenient to define the **conductance** of a circuit element as the inverse of its resistance. The symbol used to denote the conductance of an element is G, where

$$G = \frac{1}{R}\ \text{siemens (S)}, \quad \text{where } 1\ \text{S} = 1\ \text{A/V} \tag{11.13}$$

TABLE 11.1 Common Resistor Values ($^1/_8$-, $^1/_4$-, $^1/_2$-, 1-, 2-W Rating)

Ω	Code	Ω	Multiplier	kΩ	Multiplier	kΩ	Multiplier	kΩ	Multiplier
10	Brn-blk-blk	100	Brown	1.0	Red	10	Orange	100	Yellow
12	Brn-red-blk	120	Brown	1.2	Red	12	Orange	120	Yellow
15	Brn-grn-blk	150	Brown	1.5	Red	15	Orange	150	Yellow
18	Brn-gry-blk	180	Brown	1.8	Red	18	Orange	180	Yellow
22	Red-red-blk	220	Brown	2.2	Red	22	Orange	220	Yellow
27	Red-vlt-blk	270	Brown	2.7	Red	27	Orange	270	Yellow
33	Org-org-blk	330	Brown	3.3	Red	33	Orange	330	Yellow
39	Org-wht-blk	390	Brown	3.9	Red	39	Orange	390	Yellow
47	Ylw-vlt-blk	470	Brown	4.7	Red	47	Orange	470	Yellow
56	Grn-blu-blk	560	Brown	5.6	Red	56	Orange	560	Yellow
68	Blu-gry-blk	680	Brown	6.8	Red	68	Orange	680	Yellow
82	Gry-red-blk	820	Brown	8.2	Red	82	Orange	820	Yellow

black	0	blue	6
brown	1	violet	7
red	2	gray	8
orange	3	white	9
yellow	4	silver	10%
green	5	gold	5%

Resistor value = $(b_1\, b_2) \times 10^{b_3}$;
b_4 = % tolerance in actual value

FIGURE 11.9 Resistor color code.

Thus, Ohm's law can be rested in terms of conductance, as

$$I = GV \tag{11.14}$$

Ohm's law is an empirical relationship that finds widespread application in electrical engineering because of its simplicity. It is, however, only an approximation of the physics of electrically conducting materials. Typically, the linear relationship between voltage and current in electrical conductors does not apply at very high voltages and currents. Further, not all electrically conducting materials exhibit linear behavior even for small voltages and currents. It is usually true, however, that for some range of voltages and currents, most elements display a linear *i-v characteristic.*

The typical construction and the circuit symbol of the resistor are shown in Fig. 11.8. Resistors made of cylindrical sections of carbon (with resistivity $\rho = 3.5 \times 10^{-5}\ \Omega\,\text{m}$) are very common and are commercially available in a wide range of values for several power ratings (as will be explained shortly). Another commonly employed construction technique for resistors employs metal film. A common power rating for resistors used in electronic circuits (e.g., in most consumer electronic appliances such as radios and television sets) is $\frac{1}{4}$ W. Table 11.1 lists the standard values for commonly used resistors and the color code associated with these values (i.e., the common combinations of the digits $b_1b_2b_3$ as defined in Fig. 11.9. For example, if the first three color bands on a resistor show the colors red ($b_1 = 2$), violet ($b_2 = 7$), and yellow ($b_3 = 4$), the resistance value can be interpreted as follows:

$$R = 27 \times 10^4 = 270{,}000\ \Omega = 270\ \text{k}\Omega$$

In Table 11.1, the leftmost column represents the complete color code; columns to the right of it only show the third color, since this is the only one that changes. For example, a 10-Ω resistor has the code brown-black-*black*, while a 100-Ω resistor has brown-black-*brown.*

In addition to the resistance in ohms, the maximum allowable power dissipation (or **power rating**) is typically specified for commercial resistors. Exceeding this power rating leads to overheating and can cause the resistor to literally start on fire. For a resistor R, the power dissipated is given by

$$P = VI = I^2R = \frac{V^2}{R} \tag{11.15}$$

That is, the power dissipated by a resistor is proportional to the square of the current flowing through it, as well as the square of the voltage across it. The following example illustrates a common engineering application of resistive elements: the resistance strain gauge.

Example 11.1 Resistance Strain Gauges

A common application of the resistance concept to engineering measurements is the resistance **strain gauge**. Strain gauges are devices that are bonded to the surface of an object, and whose resistance varies as a function of the surface strain experienced by the object. Strain gauges may be used to perform measurements of strain, stress, force, torque, and pressure. Recall that the resistance of a cylindrical conductor of cross-sectional area A, length L, and conductivity σ is given by the expression

$$R = \frac{L}{\sigma A}$$

If the conductor is compressed or elongated as a consequence of an external force, its dimensions will change, and with them its resistance. In particular, if the conductor is stretched, its cross-sectional area will decrease and the resistance will increase. If the conductor is compressed, its resistance decreases, since the length, L, will decrease. The relationship between change in resistance and change in length is given by the gauge factor, G, defined by

$$G = \frac{\Delta R/R}{\Delta L/L}$$

and since the strain ε is defined as the fractional change in length of an object by the formula

$$\varepsilon = \frac{\Delta L}{L}$$

the change in resistance due to an applied strain ε is given by the expression

$$\Delta R = R_0 G \varepsilon$$

where R_0 is the resistance of the strain gauge under no strain and is called the zero strain resistance. The value of G for resistance strain gauges made of metal foil is usually about 2.

Figure 11.10 depicts a typical foil strain gauge. The maximum strain that can be measured by a foil gauge is about 0.4–0.5%; that is, $\Delta L/L = 0.004$ to 0.005. For a 120-Ω gauge, this corresponds to a change in resistance of the order of 0.96–1.2 Ω. Although this change in resistance is very small, it can be detected by means of suitable circuitry. Resistance strain gauges are usually connected in a circuit called the Wheatstone bridge, which we analyze later in this section.

Open and Short Circuits

Two convenient idealizations of the resistance element are provided by the limiting cases of Ohm's law as the resistance of a circuit element approaches zero or infinity. A circuit element with resistance approaching zero is called a **short circuit**. Intuitively, one would expect a short circuit to allow for unimpeded flow of current. In fact, metallic conductors (e.g., short wires of large diameter) approximate the behavior of a short circuit. Formally, a short circuit is defined as a circuit element across which the voltage is zero, regardless of the current flowing through it. Figure 11.11 depicts the circuit symbol for an ideal short circuit.

TABLE 11.2 Resistance of Copper Wire

AWG Size	Number of Strands	Diameter per Strand	Resistance per 1000 ft (Ω)
24	Solid	0.0201	28.4
24	7	0.0080	28.4
22	Solid	0.0254	18.0
22	7	0.0100	19.0
20	Solid	0.0320	11.3
20	7	0.0126	11.9
18	Solid	0.0403	7.2
18	7	0.0159	7.5
16	Solid	0.0508	4.5
16	19	0.0113	4.7

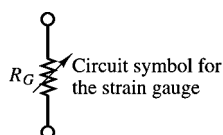



FIGURE 11.10 The resistance strain gauge.

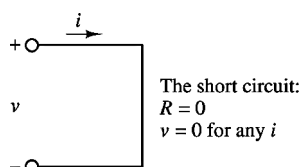



FIGURE 11.11 The short circuit.

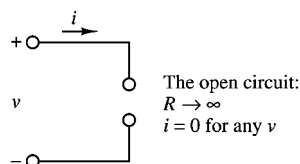



FIGURE 11.12 The open circuit.

Physically, any wire or other metallic conductor will exhibit some resistance, though small. For practical purposes, however, many elements approximate a short circuit quite accurately under certain conditions. For example, a large-diameter copper pipe is effectively a short circuit in the context of a residential electrical power supply, while in a low-power microelectronic circuit (e.g., an FM radio) a short length of 24 gauge wire (refer to Table 11.2 for the resistance of 24 gauge wire) is a more than adequate short circuit.

A circuit element whose resistance approaches infinity is called an **open circuit**. Intuitively, one would expect no current to flow through an open circuit, since it offers infinite resistance to any current. In an open circuit, we would expect to see zero current regardless of the externally applied voltage. Figure 11.12 illustrates this idea.

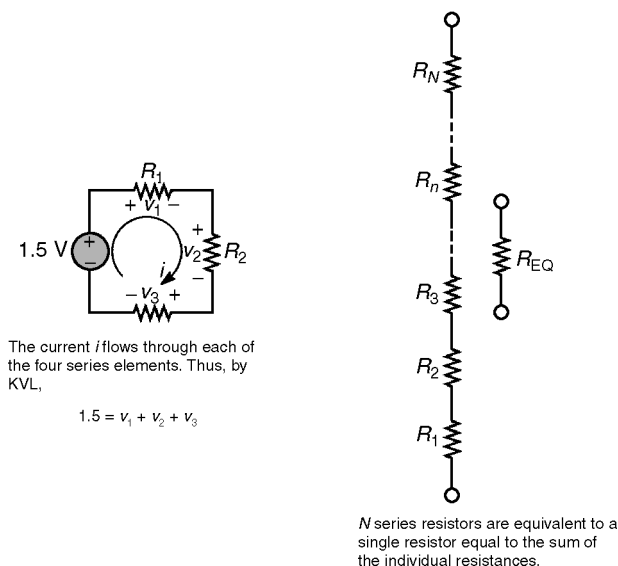



FIGURE 11.13 Voltage divider rule.

In practice, it is not too difficult to approximate an open circuit; any break in continuity in a conducting path amounts to an open circuit. The idealization of the open circuit, as defined in Fig. 11.12, does not hold, however, for very high voltages. The insulating material between two insulated terminals will break down at a sufficiently high voltage. If the insulator is air, ionized particles in the neighborhood of the two conducting elements may lead to the phenomenon of arcing; in other words, a pulse of current may be generated that momentarily jumps a gap between conductors (thanks to this principle, we are able to ignite the air-fuel mixture in a spark-ignition internal combustion engine by means of spark plugs). The ideal open and short circuits are useful concepts and find extensive use in circuit analysis.

Series Resistors and the Voltage Divider Rule

Although electrical circuits can take rather complicated forms, even the most involved circuits can be reduced to combinations of circuit elements *in parallel* and *in series*. Thus, it is important that you become acquainted with parallel and series circuits as early as possible, even before formally approaching the topic of network analysis. Parallel and series circuits have a direct relationship with Kirchhoff's laws. The objective of this section and the next is to illustrate two common circuits based on series and parallel combinations of resistors: the voltage and current dividers. These circuits form the basis of all network analysis; it is therefore important to master these topics as early as possible.

For an example of a series circuit, refer to the circuit of Fig. 11.13, where a battery has been connected to resistors R_1, R_2, and R_3. The following definition applies.

Definition

Two or more circuit elements are said to be **in series** if the same current flows through each of the elements.

The three resistors could thus be replaced by a single resistor of value R_{EQ} without changing the amount of current required of the battery. From this result we may extrapolate to the more general relationship defining the equivalent resistance of N series resistors:

$$R_{EQ} = \sum_{n=1}^{N} R_n \qquad (11.16)$$

which is also illustrated in Fig. 11.13. A concept very closely tied to series resistors is that of the **voltage divider**.

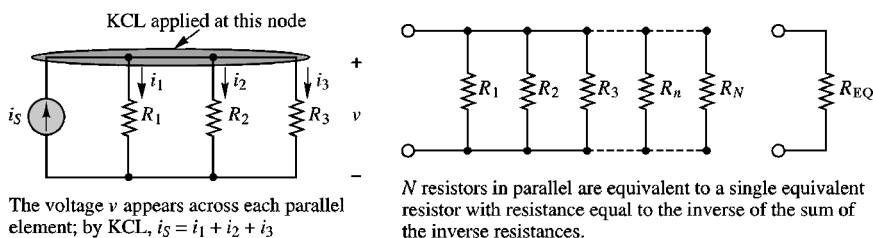



FIGURE 11.14 Parallel circuits.

The general form of the voltage divider rule for a circuit with N series resistors and a voltage source is

$$v_n = \frac{R_n}{R_1 + R_2 + \cdots + R_n + \cdots + R_N} v_S \tag{11.17}$$

Parallel Resistors and the Current Divider Rule

A concept analogous to that of the voltage may be developed by applying Kirchhoff's current law to a circuit containing only parallel resistances.

Definition

Two or more circuit elements are said to be **in parallel** if the same voltage appears across each of the elements. (See Fig. 11.14.)

N resistors in parallel act as a single equivalent resistance, R_{EQ}, given by the expression

$$\frac{1}{R_{EQ}} = \frac{1}{R_1} + \frac{1}{R_2} + \cdots + \frac{1}{R_N} \tag{11.18}$$

or

$$R_{EQ} = \frac{1}{1/R_1 + 1/R_2 + \cdots + 1/R_N} \tag{11.19}$$

Very often in the remainder of this book we shall refer to the parallel combination of two or more resistors with the following notation:

$$R_1 \parallel R_2 \parallel \cdots$$

where the symbol $\parallel$ signifies "in parallel with."

The general expression for the current divider for a circuit with N parallel resistors is the following:

$$i_n = \frac{1/R_n}{1/R_1 + 1/R_2 + \cdots + 1/R_n + \cdots + 1/R_N} i_S \quad \text{Current divider} \tag{11.20}$$

Example 11.2 The Wheatstone Bridge

The **Wheatstone bridge** is a resistive circuit that is frequently encountered in a variety of measurement circuits. The general form of the bridge is shown in Fig. 11.15(a), where R_1, R_2, and R_3 are known, while R_x is an unknown resistance, to be determined. The circuit may also be redrawn as shown in Fig. 11.15(b). The latter circuit will be used to demonstrate the use of the voltage divider rule in a mixed series-parallel circuit.

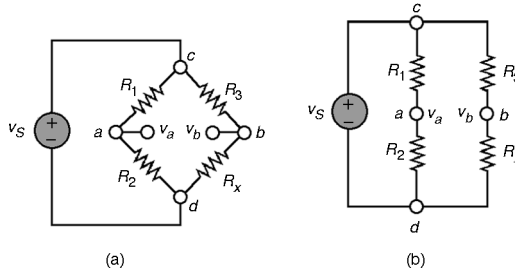



FIGURE 11.15 Wheatstone bridge circuits.

The objective is to determine the unknown resistance R_x.

1. Find the value of the voltage $v_{ad} = v_{ad} - v_{bd}$ in terms of the four resistances and the source voltage, v_S. Note that since the reference point d is the same for both voltages, we can also write $v_{ab} = v_a - v_b$.
2. If $R_1 = R_2 = R_3 = 1\ \text{k}\Omega$, $v_S = 12$ V, and $v_{ab} = 12$ mV, what is the value of R_x?

Solution

1. First, we observe that the circuit consists of the parallel combination of three subcircuits: the voltage source, the series combination of R_1 and R_2, and the series combination of R_3 and R_x. Since these three subcircuits are in parallel, the same voltage will appear across each of them, namely, the source voltage, v_S.
   Thus, the source voltage divides between each resistor pair, R_1-R_2 and R_3-R_x, according to the voltage divider rule: v_a is the fraction of the source voltage appearing across R_2, while v_b is the voltage appearing across R_x:

$$v_a = v_S \frac{R_2}{R_1 + R_2} \quad \text{and} \quad v_b = v_S \frac{R_x}{R_3 + R_x}$$

   Finally, the voltage difference between points a and b is given by

$$v_{ab} = v_a - v_b = v_S \left(\frac{R_2}{R_1 + R_2} - \frac{R_x}{R_3 + R_x} \right)$$

   This result is very useful and quite general, and it finds application in numerous practical circuits.
2. In order to solve for the unknown resistance, we substitute the numerical values in the preceding equation to obtain

$$0.012 = 12\left(\frac{1000}{2000} - \frac{R_x}{1000 + R_x} \right)$$

   which may be solved for R_x to yield

$$R_x = 996\ \Omega$$

Practical Voltage and Current Sources

Idealized models of voltage and current sources fail to take into consideration the finite-energy nature of practical voltage and current sources. The objective of this section is to extend the ideal models to models that are capable of describing the physical limitations of the voltage and current sources used in practice. Consider, for example, the model of an ideal voltage source. As the load resistance (R) decreases, the source is required to provide increasing amounts of current to maintain the voltage $v_S(t)$ across

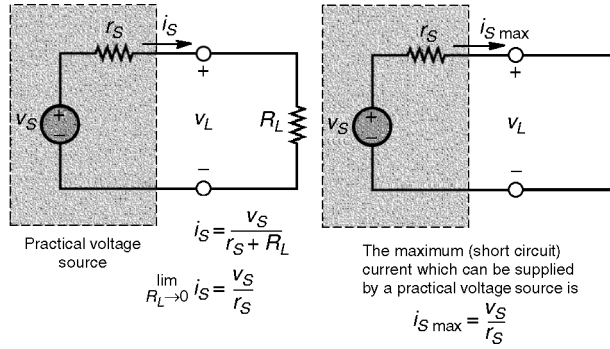



FIGURE 11.16 Practical voltage source.

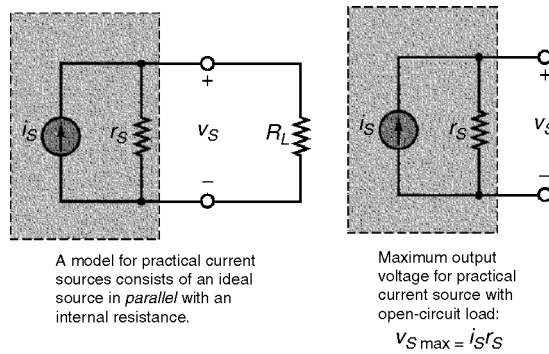



FIGURE 11.17 Practical current source.

its terminal:

$$i(t) = \frac{v_S(t)}{R} \tag{11.21}$$

This circuit suggests that the ideal voltage source is required to provide an infinite amount of current to the load, in the limit as the load resistance approaches zero.

Figure 11.16 depicts a model for a practical voltage source; this is composed of an ideal voltage source, v_S, in series with a resistance, r_S. The resistance r_S in effect poses a limit to the maximum current the voltage source can provide:

$$i_{S\ \max} = \frac{v_S}{r_S} \tag{11.22}$$

It should be apparent that a desirable feature of an ideal voltage source is a very small internal resistance, so that the current requirements of an arbitrary load may be satisfied.

A similar modification of the ideal current source model is useful to describe the behavior of a practical current source. The circuit illustrated in Fig. 11.17 depicts a simple representation of a practical current source, consisting of an ideal source in parallel with a resistor. Note that as the load resistance approaches infinity (i.e., an open circuit), the output voltage of the current source approaches its limit,

$$v_{S\ \max} = i_S r_S \tag{11.23}$$

A good current source should be able to approximate the behavior of an ideal current source. Therefore, a desirable characteristic for the internal resistance of a current source is that it be as large as possible.

Measuring Devices

The Ammeter

The **ammeter** is a device that, when connected in series with a circuit element, can measure the current flowing through the element. Figure 11.18 illustrates this idea. From Fig. 11.18, two requirements are evident for obtaining a correct measurement of current:

1. The ammeter must be placed in series with the element whose current is to be measured (e.g., resistor R_2).
2. The ammeter should not resist the flow of current (i.e., cause a voltage drop), or else it will not be measuring the true current flowing the circuit. *An ideal ammeter has zero internal resistance.*

The Voltmeter

The **voltmeter** is a device that can measure the voltage across a circuit element. Since voltage is the difference in potential between two points in a circuit, the voltmeter needs to be connected across the element whose voltage we wish to measure. A voltmeter must also fulfill two requirements:

1. The voltmeter must be placed in parallel with the element whose voltage it is measuring.
2. The voltmeter should draw no current away from the element whose voltage it is measuring, or else it will not be measuring the true voltage across that element. Thus, *an ideal voltmeter has infinite internal resistance.*

Figure 11.19 illustrates these two points.

Once again, the definitions just stated for the ideal voltmeter and ammeter need to be augmented by considering the practical limitations of the devices. A practical ammeter will contribute some series resistance to the circuit in which it is measuring current; a practical voltmeter will not act as an ideal open circuit but will always draw some current from the measured circuit. Figure 11.20 depicts the circuit models for the practical ammeter and voltmeter.

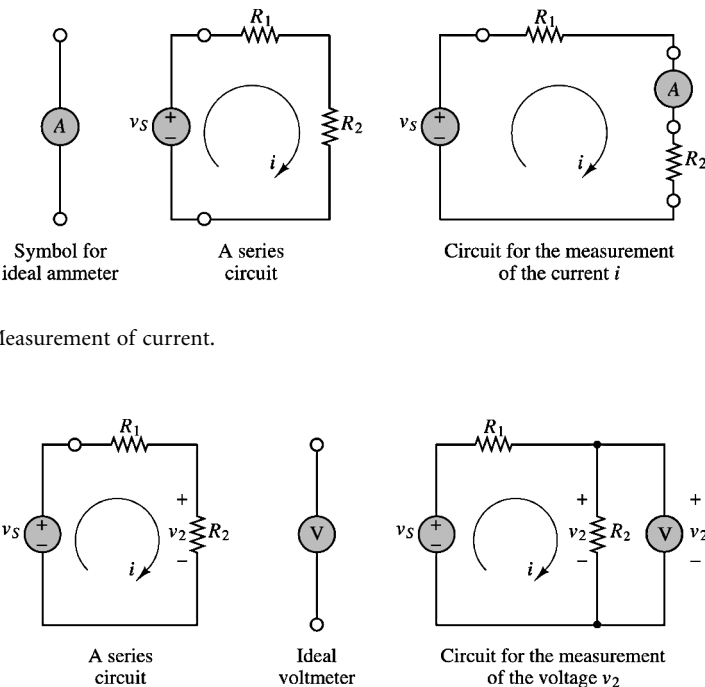



FIGURE 11.18 Measurement of current.

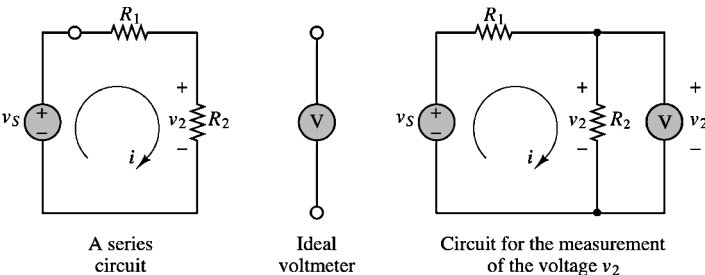



FIGURE 11.19 Measurement of voltage.

FIGURE 11.20 Models for practical ammeter and voltmeter.

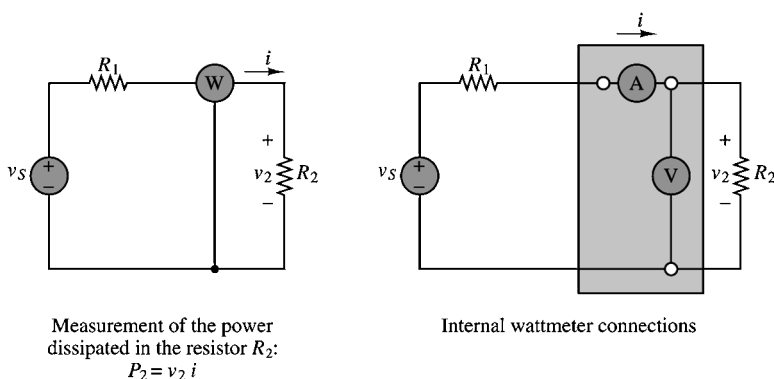
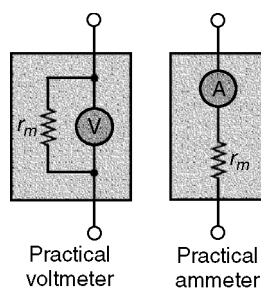



FIGURE 11.21 Measurement of power.

All of the considerations that pertain to practical ammeters and voltmeters can be applied to the operation of a **wattmeter**, a measuring instrument that provides a measurement of the power dissipated by a circuit element, since the wattmeter is in effect made up of a combination of a voltmeter and an ammeter.

Figure 11.21 depicts the typical connection of a wattmeter in the same series circuit used in the preceding paragraphs. In effect, the wattmeter measures the current flowing through the load and, simultaneously, the voltage across it multiplies the two to provide a reading of the power dissipated by the load.

11.3 Resistive Network Analysis

This section will illustrate the fundamental techniques for the analysis of resistive circuits. The methods introduced are based on Kirchhoff's and Ohm's laws. The main thrust of the section is to introduce and illustrate various methods of circuit analysis that will be applied throughout the book.

The Node Voltage Method

Node voltage analysis is the most general method for the analysis of electrical circuits. In this section, its application to linear resistive circuits will be illustrated. The **node voltage method** is based on defining the voltage at each node as an independent variable. One of the nodes is selected as a **reference node** (usually—but not necessarily—ground), and each of the other node voltages is referenced to this node. Once each node voltage is defined, Ohm's law may be applied between any two adjacent nodes in order to determine the current flowing in each branch. In the node voltage method, *each branch current is expressed in terms of one or more node voltages*; thus, currents do not explicitly enter into the equations. Figure 11.22 illustrates how one defines branch currents in this method.

In the node voltage method, we assign the node voltages v_a and v_b; the branch current flowing from a to b is then expressed in terms of these node voltages.

$$i = \frac{v_a - v_b}{R}$$

v_a — R — v_b, i

FIGURE 11.22 Branch current formulation in nodal analysis.

By KCL: $i_1 = i_2 + i_3$. In the node voltage method, we express KCL by

$$\frac{v_a - v_b}{R_1} = \frac{v_b - v_c}{R_2} + \frac{v_b - v_d}{R_3}$$

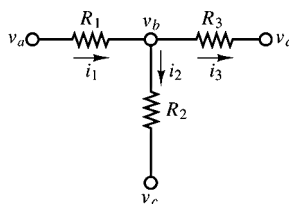



FIGURE 11.23 Use of KCL in nodal analysis.

Once each branch current is defined in terms of the node voltages, Kirchhoff's current law is applied at each node. The particular form of KCL employed in the nodal analysis equates the sum of the currents into the node to the sum of the currents leaving the node:

$$\sum i_{\text{in}} = \sum i_{\text{out}} \tag{11.24}$$

Figure 11.23 illustrates this procedure.

The systematic application of this method to a circuit with n nodes would lead to writing n linear equations. However, one of the node voltages is the reference voltage and is therefore already known, since it is usually assumed to be zero. Thus, we can write $n - 1$ *independent linear equations* in the $n - 1$ independent variables (the node voltages). Nodal analysis provides the minimum number of equations required to solve the circuit, since any branch voltage or current may be determined from knowledge of nodal voltages.

The nodal analysis method may also be defined as a sequence of steps, as outlined below.

Node Voltage Analysis Method

1. Select a reference node (usually ground). All other node voltages will be referenced to this node.
2. Define the remaining $n - 1$ node voltages as the independent variables.
3. Apply KCL at each of the $n - 1$ nodes, expressing each current in terms of the adjacent node voltages.
4. Solve the linear system of $n - 1$ equations in $n - 1$ unknowns.

In a circuit containing n nodes we can write at most $n - 1$ independent equations.

The Mesh Current Method

In the mesh current method, we observe that a current flowing through a resistor in a specified direction defines the polarity of the voltage across the resistor, as illustrated in Fig. 11.24, and that the sum of the voltages around a closed circuit must equal zero, by KVL. Once a convention is established regarding the direction of current flow around a mesh, simple application of KVL provides the desired equation. Figure 11.25 illustrates this point.

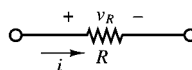



FIGURE 11.24 Basic principle of mesh analysis.

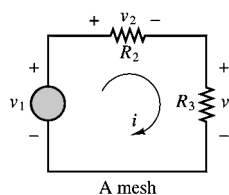



FIGURE 11.25 Use of KVL in mesh analysis.

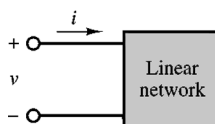



FIGURE 11.26 One-port network.

The number of equations one obtains by this technique is equal to the number of meshes in the circuit. All branch currents and voltages may subsequently be obtained from the mesh currents, as will presently be shown. Since meshes are easily identified in a circuit, this method provides a very efficient and systematic procedure for the analysis of electrical circuits. The following section outlines the procedure used in applying the mesh current method to a linear circuit.

Mesh Current Analysis Method

1. Define each mesh current consistently. We shall always define mesh currents clockwise, for convenience.
2. Apply KVL around each mesh, expressing each voltage in terms of one or more mesh currents.
3. Solve the resulting linear system of equations with mesh currents as the independent variables.

In mesh analysis, it is important to be consistent in choosing the direction of current flow. To avoid confusion in writing the circuit equations, mesh currents will be defined exclusively clockwise when we are using this method.

One-Port Networks and Equivalent Circuits

This general circuit representation is shown in Fig. 11.26. This configuration is called a **one-port network** and is particularly useful for introducing the notion of equivalent circuits. Note that the network of Fig. 11.26 is completely described by its i-v characteristic.

Thévenin and Norton Equivalent Circuits

This section discusses one of the most important topics in the analysis of electrical circuits: the concept of an **equivalent circuit**. It will be shown that it is always possible to view even a very complicated circuit in terms of much simpler *equivalent* source and load circuits, and that the transformations leading to equivalent circuits are easily managed, with a little practice. In studying node voltage and mesh current analysis, you may have observed that there is a certain correspondence (called **duality**) between current sources and voltage sources, on the one hand, and parallel and series circuits, on the other. This duality appears again very clearly in the analysis of equivalent circuits: it will shortly be shown that equivalent circuits fall into one of two classes, involving either voltage or current sources and (respectively) either

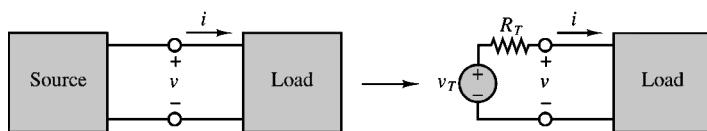



FIGURE 11.27 Illustration of Thévenin theorem.

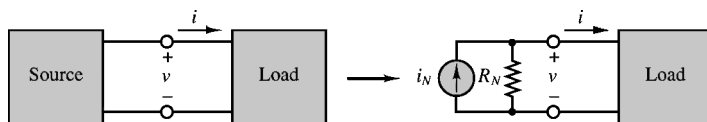



FIGURE 11.28 Illustration of Norton theorem.

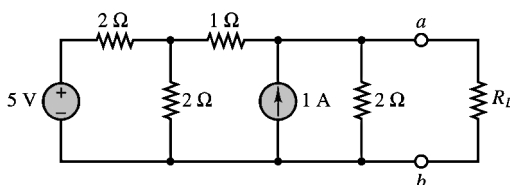



FIGURE 11.29 Computation of Thévenin resistance.

series or parallel resistors, reflecting this same principle of duality. The discussion of equivalent circuits begins with the statement of two very important theorems, summarized in Figs. 11.27 and 11.28.

The Thévenin Theorem

As far as a load is concerned, any network composed of ideal voltage and current sources, and of linear resistors, may be represented by an equivalent circuit consisting of an ideal voltage source, v_T, in series with an equivalent resistance, R_T.

The Norton Theorem

As far as a load is concerned, any network composed of ideal voltage and current sources, and of linear resistors, may be represented by an equivalent circuit consisting of an ideal current source, i_N, in parallel with an equivalent resistance, R_N.

Determination of Norton or Thévenin Equivalent Resistance

The first step in computing a Thévenin or Norton equivalent circuit consists of finding the equivalent resistance presented by the circuit at its terminals. This is done by setting all sources in the circuit equal to zero and computing the effective resistance between terminals. The voltage and current sources present in the circuit are set to zero as follows: voltage sources are replaced by short circuits, current sources by open circuits. We can produce a set of simple rules as an aid in the computation of the Thévenin (or Norton) equivalent resistance for a linear resistive circuit.

Computation of Equivalent Resistance of a One-Port Network:

1. Remove the load.
2. Zero all voltage and current sources
3. Compute the total resistance between load terminals, *with the load removed.* This resistance is equivalent to that which would be encountered by a current source connected to the circuit in place of the load.

For example, the equivalent resistance of the circuit of Fig. 11.29 as seen by the load is:

$$R_{eq} = ((2 \| 2) + 1) \| 2 = 1 \ \Omega.$$

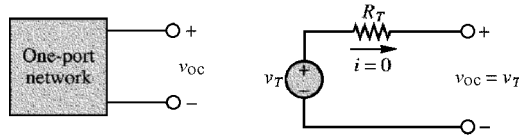



FIGURE 11.30 Equivalence of open-circuit and Thévenin voltage.

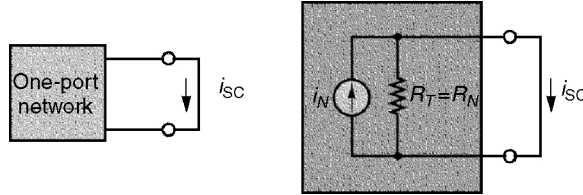



FIGURE 11.31 Illustration of Norton equivalent circuit.

Computing the Thévenin Voltage

The Thévenin equivalent voltage is defined as follows: the equivalent (Thévenin) source voltage is equal to the **open-circuit voltage** present at the load terminals with the load removed.

This states that in order to compute v_T, it is sufficient to remove the load and to compute the open-circuit voltage at the one-port terminals. Figure 11.30 illustrates that the open-circuit voltage, v_{OC}, and the Thévenin voltage, v_T, must be the same if the Thévenin theorem is to hold. This is true because in the circuit consisting of v_T and R_T, the voltage v_{OC} must equal v_T, since no current flows through R_T and therefore the voltage across R_T is zero. Kirchhoff's voltage law confirms that

$$v_T = R_T(0) + v_{OC} = v_{OC} \tag{11.25}$$

Computing the Norton Current

The computation of the Norton equivalent current is very similar in concept to that of the Thévenin voltage. The following definition will serve as a starting point.

Definition

The Norton equivalent current is equal to the **short-circuit current** that would flow were the load replaced by a short circuit.

An explanation for the definition of the Norton current is easily found by considering, again, an arbitrary one-port network, as shown in Fig. 11.31, where the one-port network is shown together with its Norton equivalent circuit.

It should be clear that the current, i_{SC}, flowing through the short circuit replacing the load is exactly the Norton current, i_N, since all of the source current in the circuit of Fig. 11.31 must flow through the short circuit.

Experimental Determination of Thévenin and Norton Equivalents

Figure 11.32 illustrates the measurement of the open-circuit voltage and short-circuit current for an arbitrary network connected to any load and also illustrates that the procedure requires some special attention, because of the nonideal nature of any practical measuring instrument. The figure clearly illustrates that in the presence of finite meter resistance, r_m, one must take this quantity into account in the computation of the short-circuit current and open-circuit voltage; v_{OC} and i_{SC} appear between quotation marks in the figure specifically to illustrate that the measured "open-circuit voltage" and "short-circuit current" are, in fact, affected by the internal resistance of the measuring instrument and are not the true quantities.

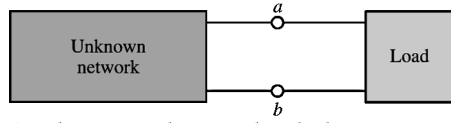
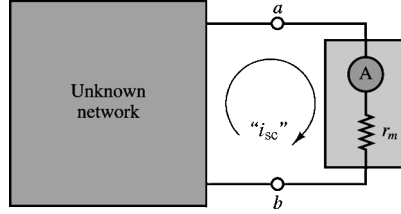
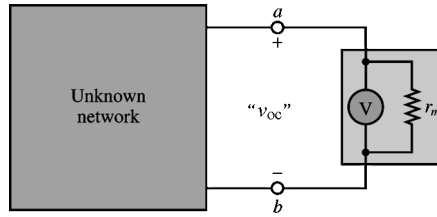



FIGURE 11.32 Measurement of open-circuit voltage and short-circuit current.

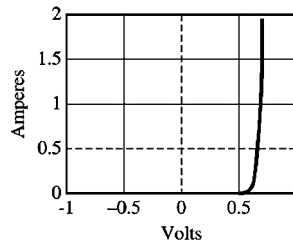



FIGURE 11.33 *i-v* characteristic of exponential resistor.

The following are expressions for the true short-circuit current and open-circuit voltage.

$$i_N = i_{SC}\left(1 + \frac{r_m}{R_T}\right)$$
$$v_T = v_{OC}\left(1 + \frac{R_T}{r_m}\right) \tag{11.26}$$

where i_N is the ideal Norton current, v_T the Thévenin voltage, and R_T the true Thévenin resistance.

Nonlinear Circuit Elements

Description of Nonlinear Elements

There are a number of useful cases in which a simple functional relationship exists between voltage and current in a nonlinear circuit element. For example, Fig. 11.33 depicts an element with an exponential *i-v* characteristic, described by the following equations:

$$i = I_0 e^{\alpha v}, \quad v > 0$$
$$i = -I_0, \quad v \le 0 \tag{11.27}$$

FIGURE 11.34 Representation of nonlinear element in a linear circuit.

There exists, in fact, a circuit element (the semiconductor diode) that very nearly satisfies this simple relationship. The difficulty in the *i*-*v* relationship of Eq. (11.27) is that it is not possible, in general, to obtain a closed-form analytical solution, even for a very simple circuit.

One approach to analyzing a circuit containing a nonlinear element might be to treat the nonlinear element as a load, and to compute the Thévenin equivalent of the remaining circuit, as shown in Fig. 11.34. Applying KVL, the following equation may then be obtained:

$$v_T = R_T i_x + v_x \tag{11.28}$$

To obtain the second equation needed to solve for both the unknown voltage, v_x, and the unknown current, i_x, it is necessary to resort to the *i*-*v* description of the nonlinear element, namely, Eq. (11.27). If, for the moment, only positive voltages are considered, the circuit is completely described by the following system:

$$\begin{aligned} i_x &= I_0 e^{\alpha v_x}, \quad v > 0 \\ v_T &= R_T i_x + v_x \end{aligned} \tag{11.29}$$

The two parts of Eq. (11.29) represent a system of two equations in two unknowns. Any numerical method of choice may now be applied to solve the system of Eqs. (11.29).

11.4 AC Network Analysis

In this section we introduce energy-storage elements, dynamic circuits, and the analysis of circuits excited by sinusoidal voltages and currents. Sinusoidal (or AC) signals constitute the most important class of signals in the analysis of electrical circuits. The simplest reason is that virtually all of the electric power used in households and industries comes in the form of sinusoidal voltages and currents.

Energy-Storage (Dynamic) Circuit Elements

The ideal resistor was introduced through Ohm's law in Section 11.2 as a useful idealization of many practical electrical devices. However, in addition to resistance to the flow of electric current, which is purely a dissipative (i.e., an energy-loss) phenomenon, electric devices may also exhibit energy-storage properties, much in the same way a spring or a flywheel can store mechanical energy. Two distinct mechanisms for energy storage exist in electric circuits: **capacitance** and **inductance**, both of which lead to the storage of energy in an electromagnetic field.

The Ideal Capacitor

A physical capacitor is a device that can store energy in the form of a charge separation when appropriately polarized by an electric field (i.e., a voltage). The simplest capacitor configuration consists of two parallel

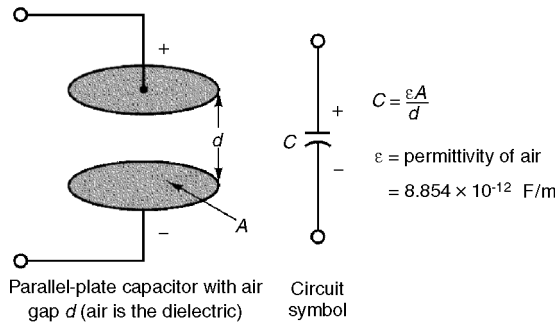



FIGURE 11.35 Structure of parallel-plate capacitor.

conducting plates of cross-sectional area A, separated by air (or another **dielectric*** material, such as mica or Teflon). Figure 11.35 depicts a typical configuration and the circuit symbol for a capacitor.

The presence of an insulating material between the conducting plates does not allow for the flow of DC current; thus, *a capacitor acts as an open circuit in the presence of DC currents*. However, if the voltage present at the capacitor terminals changes as a function of time, so will the charge that has accumulated at the two capacitor plates, since the degree of polarization is a function of the applied electric field, which is time-varying. In a capacitor, the charge separation caused by the polarization of the dielectric is proportional to the external voltage, that is, to the applied electric field:

$$Q = CV \tag{11.30}$$

where the parameter C is called the *capacitance* of the element and is a measure of the ability of the device to accumulate, or store, charge. The unit of capacitance is the coulomb/volt and is called the **farad** (F). The farad is an unpractically large unit; therefore, it is common to use microfarads (1 μF = 10^{-6} F) or picofarads (1 pF = 10^{-12} F). From Eq. (11.30) it becomes apparent that if the external voltage applied to the capacitor plates changes in time, so will the charge that is internally stored by the capacitor:

$$q(t) = Cv(t) \tag{11.31}$$

Thus, although no current can flow through a capacitor if the voltage across it is constant, a time-varying voltage will cause charge to vary in time. The change with time in the stored charge is analogous to a current. The relationship between the current and voltage in a capacitor is as follows:

$$i(t) = C\frac{dv(t)}{dt} \tag{11.32}$$

If the above differential equation is integrated, one can obtain the following relationship for the voltage across a capacitor:

$$v_C(t) = \frac{1}{C}\int_{-\infty}^{t_0} i_C\, dt \tag{11.33}$$

Equation (11.33) indicates that the capacitor voltage depends on the past current through the capacitor, up until the present time, t. Of course, one does not usually have precise information regarding the flow

* A dielectric material contains a large number of electric dipoles, which become polarized in the presence of an electric field.

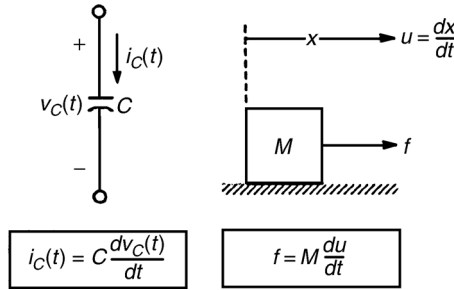



FIGURE 11.36 Defining equation for the ideal capacitor, and analogy with force-mass system.

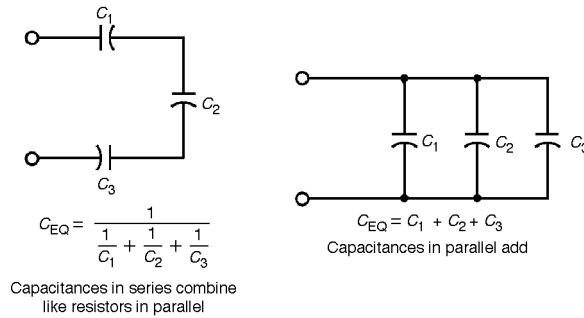



FIGURE 11.37 Combining capacitors in a circuit.

of capacitor current for all past time, and so it is useful to define the initial voltage (or *initial condition*) for the capacitor according to the following, where t_0 is an arbitrary initial time:

$$V_0 = v_C(t = t_0) = \frac{1}{C}\int_{-\infty}^{t} i_C\,dt \tag{11.34}$$

The capacitor voltage is now given by the expression

$$v_C(t) = \frac{1}{C}\int_{t_0}^{t} i_C\,dt + V_0 \quad t \geq t_0 \tag{11.35}$$

The significance of the initial voltage, V_0, is simply that at time t_0 some charge is stored in the capacitor, giving rise to a voltage, $v_C(t_0)$, according to the relationship $Q = CV$. Knowledge of this initial condition is sufficient to account for the entire past history of the capacitor current. (See Fig. 11.36.)

From the standpoint of circuit analysis, it is important to point out that capacitors connected in series and parallel can be combined to yield a single equivalent capacitance. The rule of thumb, which is illustrated in Fig. 11.37, is the following: capacitors in parallel add; capacitors in series combine according to the same rules used for resistors connected in parallel.

Physical capacitors are rarely constructed of two parallel plates separated by air, because this configuration yields very low values of capacitance, unless one is willing to tolerate very large plate areas. In order to increase the capacitance (i.e., the ability to store energy), physical capacitors are often made of tightly rolled sheets of metal film, with a dielectric (paper or Mylar) sandwiched in-between. Table 11.3 illustrates typical values, materials, maximum voltage ratings, and useful frequency ranges for various

TABLE 11.3 Capacitors

Material	Capacitance Range	Maximum Voltage (V)	Frequency Range (Hz)
Mica	1 pF to 0.1 μF	100–600	10^3–10^{10}
Ceramic	10 pF to 1 μF	50–1000	10^3–10^{10}
Mylar	0.001 to 10 μF	50–500	10^2–10^8
Paper	1000 pF to 50 μF	100–105	10^2–10^8
Electrolytic	0.1 μF to 0.2 F	3–600	10–10^4

types of capacitors. The voltage rating is particularly important, because any insulator will break down if a sufficiently high voltage is applied across it. The energy stored in a capacitor is given by

$$W_C(t) = \frac{1}{2}Cv_C^2(t) \text{ (J)}$$

Example 11.3 Capacitive Displacement Transducer and Microphone

As shown in Fig. 11.26, the capacitance of a parallel-plate capacitor is given by the expression

$$C = \frac{\varepsilon A}{d}$$

where ε is the **permittivity** of the dielectric material, A the area of each of the plates, and d their separation. The permittivity of air is $\varepsilon_0 = 8.854 \times 10^{-12}$ F/m, so that two parallel plates of area 1 m^2, separated by a distance of 1 mm, would give rise to a capacitance of 8.854×10^{-3} μF, a very small value for a very large plate area. This relative inefficiency makes parallel-plate capacitors impractical for use in electronic circuits. On the other hand, parallel-plate capacitors find application as *motion transducers*, that is, as devices that can measure the motion or displacement of an object. In a capacitive motion transducer, the air gap between the plates is designed to be variable, typically by fixing one plate and connecting the other to an object in motion. Using the capacitance value just derived for a parallel-plate capacitor, one can obtain the expression

$$C = \frac{8.854 \times 10^{-3} A}{x}$$

where C is the capacitance in picofarad, A is the area of the plates in square millimeter, and x is the (variable) distance in milimeter. It is important to observe that the change in capacitance caused by the displacement of one of the plates is nonlinear, since the capacitance varies as the inverse of the displacement. For small displacements, however, the capacitance varies approximately in a linear fashion.

The *sensitivity, S,* of this motion transducer is defined as the slope of the change in capacitance per change in displacement, x, according to the relation

$$S = \frac{dC}{dx} = -\frac{8.854 \times 10^{-3} A}{2x^2} \text{ (pF/mm)}$$

Thus, the sensitivity increases for small displacements. This behavior can be verified by plotting the capacitance as a function of x and noting that as x approaches zero, the slope of the nonlinear $C(x)$ curve becomes steeper (thus the greater sensitivity). Figure 11.38 depicts this behavior for a transducer with area equal to 10 mm^2.

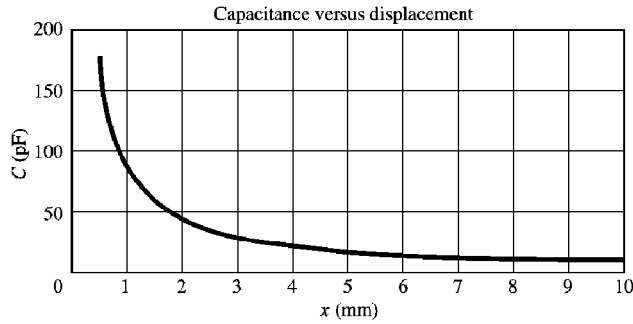



FIGURE 11.38 Response of a capacitive displacement transducer.

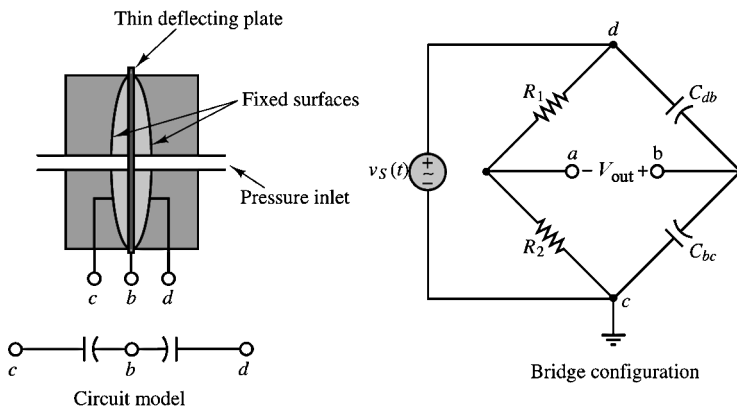



FIGURE 11.39 Capacitive pressure transducer and related bridge circuit.

This simple capacitive displacement transducer actually finds use in the popular *capacitive (or condenser) microphone*, in which the sound pressure waves act to displace one of the capacitor plates. The change in capacitance can then be converted into a change in voltage or current by means of a suitable circuit. An extension of this concept that permits measurement of differential pressures is shown in simplified form in Fig. 11.39. In the figure, a three-terminal variable capacitor is shown to be made up of two fixed surfaces (typically, spherical depressions ground into glass disks and coated with a conducting material) and of a deflecting plate (typically made of steel) sandwiched between the glass disks. Pressure inlet orifices are provided, so that the deflecting plate can come into contact with the fluid whose pressure it is measuring. When the pressure on both sides of the deflecting plate is the same, the capacitance between terminals b and d, C_{bd}, will be equal to that between terminals b and c, C_{bc}. If any pressure differential exists, the two capacitances will change, with an increase on the side where the deflecting plate has come closer to the fixed surface and a corresponding decrease on the other side.

This behavior is ideally suited for the application of a bridge circuit, similar to the Wheatstone bridge circuit illustrated in Example 11.2, and also shown in Fig. 11.39. In the bridge circuit, the output voltage, v_{out}, is precisely balanced when the differential pressure across the transducer is zero, but it will deviate from zero whenever the two capacitances are not identical because of a pressure differential across the transducer. We shall analyze the bridge circuit later in Example 11.4.

The Ideal Inductor

The ideal inductor is an element that has the ability to store energy in a magnetic field. Inductors are typically made by winding a coil of wire around a core, which can be an insulator or a ferromagnetic material, shown in Fig. 11.40. When a current flows through the coil, a magnetic field is established, as

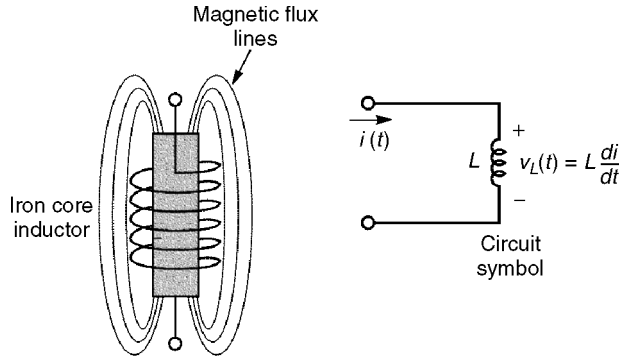



FIGURE 11.40 Iron-core inductor.

you may recall from early physics experiments with electromagnets. In an ideal inductor, the resistance of the wire is zero, so that a constant current through the inductor will flow freely without causing a voltage drop. In other words, *the ideal inductor acts as a short circuit in the presence of DC currents*. If a time-varying voltage is established across the inductor, a corresponding current will result, according to the following relationship:

$$v_L(t) = L\frac{di_L}{dt} \tag{11.36}$$

where L is called the *inductance* of the coil and is measured in henry (H), where

$$1 \text{ H} = 1 \text{ V sec/A} \tag{11.37}$$

Henrys are reasonable units for practical inductors; millihenrys (mH) and microhenrys (μH) are also used.

The inductor current is found by integrating the voltage across the inductor:

$$i_L(t) = \frac{1}{L}\int_{-\infty}^{t} v_L \, dt \tag{11.38}$$

If the current flowing through the inductor at time $t = t_0$ is known to be I_0, with

$$I_0 = i_L(t = t_0) = \frac{1}{L}\int_{-\infty}^{t_0} v_L \, dt \tag{11.39}$$

then the inductor current can be found according to the equation

$$i_L(t) = \frac{1}{L}\int_{t_0}^{t} v_L \, dt + I_0 \quad t \geq t_0 \tag{11.40}$$

Inductors in series add. Inductors in parallel combine according to the same rules used for resistors connected in parallel. See Figs. 11.41–11.43.

Table 11.4 and Figs. 11.36, 11.41, and 11.43 illustrate a useful analogy between ideal electrical and mechanical elements.

TABLE 11.4 Analogy Between Electrical and Mechanical Variables

Mechanical System	Electrical System
Force, f (N)	Current, i (A)
Velocity, μ (m/sec)	Voltage, v (V)
Damping, B (N sec/m)	Conductance, $1/R$ (S)
Compliance, 1/k (m/N)	Inductance, L (H)
Mass, M (kg)	Capacitance, C (F)

The defining equation for the inductance circuit element is analogous to the equation of motion of a spring acted upon by a force.

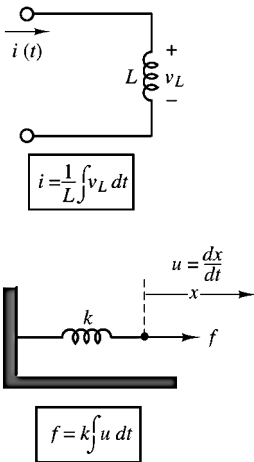



FIGURE 11.41 Defining equation for the ideal inductor and analogy with force-spring system.

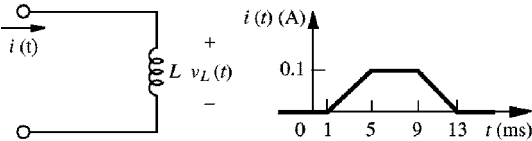



FIGURE 11.42 Combining inductors in a circuit.

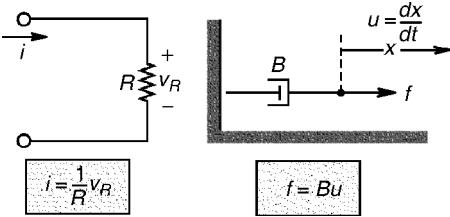



FIGURE 11.43 Analogy between electrical and mechanical elements.

Time-Dependent Signal Sources

Figure 11.44 illustrates the convention that will be employed to denote time-dependent signal sources.

One of the most important classes of time-dependent signals is that of **periodic signals**. These signals appear frequently in practical applications and are a useful approximation of many physical phenomena. A periodic signal $x(t)$ is a signal that satisfies the following equation:

$$x(t) = x(t + nT) \quad n = 1, 2, 3, \ldots \tag{11.41}$$

where T is the **period** of $x(t)$. Figure 11.45 illustrates a number of the periodic waveforms that are typically encountered in the study of electrical circuits. Waveforms such as the sine, triangle, square, pulse, and sawtooth waves are provided in the form of voltages (or, less frequently, currents) by commercially available **signal** (or **waveform**) **generators**. Such instruments allow for selection of the waveform peak amplitude, and of its period.

As stated in the introduction, sinusoidal waveforms constitute by far the most important class of time-dependent signals. Figure 11.46 depicts the relevant parameters of a sinusoidal waveform. A generalized sinusoid is defined as follows:

$$x(t) = A \cos(\omega t + \phi) \tag{11.42}$$

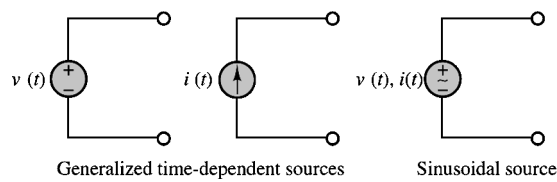



FIGURE 11.44 Time-dependent signal sources.

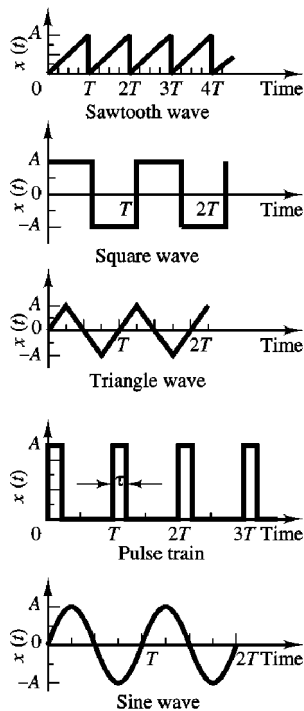



FIGURE 11.45 Periodic signal waveforms.

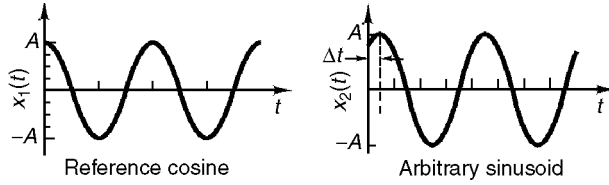



FIGURE 11.46 Sinusoidal waveforms.

where A is the **amplitude**, ω the **radian frequency**, and ϕ the **phase**. Figure 11.46 summarizes the definitions of A, ω, and ϕ for the waveforms

$$x_1(t) = A\cos(\omega t) \quad \text{and} \quad x_2(t) = A\cos(\omega t + \phi)$$

where

$$\begin{aligned} f &= \text{natural frequency} = \frac{1}{T} \text{ (cycles/sec, or Hz)} \\ \omega &= \text{radian frequency} = 2\pi f \text{ (radians/sec)} \\ \phi &= 2\pi\frac{\Delta T}{T} \text{ (radians)} = 360\frac{\Delta T}{T} \text{ (degrees)} \end{aligned} \tag{11.43}$$

The phase shift, ϕ, permits the representation of an arbitrary sinusoidal signal. Thus, the choice of the reference cosine function to represent sinusoidal signals—arbitrary as it may appear at first—does not restrict the ability to represent all sinusoids. For example, one can represent a sine wave in terms of a cosine wave simply by introducing a phase shift of $\pi/2$ radians:

$$A\sin(\omega t) = A\cos\left(\omega t - \frac{\pi}{2}\right) \tag{11.44}$$

It is important to note that, although one usually employs the variable ω (in units of radians per second) to denote sinusoidal frequency, it is common to refer to natural frequency, f, in units of cycles per second, or hertz (Hz). The relationship between the two is the following:

$$\omega = 2\pi f \tag{11.45}$$

Average and RMS Values

Now that a number of different signal waveforms have been defined, it is appropriate to define suitable measurements for quantifying the strength of a time-varying electrical signal. The most common types of measurements are the **average** (or **DC**) **value** of a signal waveform, which corresponds to just measuring the mean voltage or current over a period of time, and the **root-mean-square** (**rms**) **value**, which takes into account the fluctuations of the signal about its average value. Formally, the operation of computing the average value of a signal corresponds to integrating the signal waveform over some (presumably, suitably chosen) period of time. We define the time-averaged value of a signal $x(t)$ as

$$\langle x(t)\rangle = \frac{1}{T}\int_0^T x(t)\,dt \tag{11.46}$$

where T is the period of integration. Figure 11.47 illustrates how this process does, in fact, correspond to computing the average amplitude of $x(t)$ over a period of T seconds.

$$\langle A\cos(\omega t + \phi)\rangle = 0$$

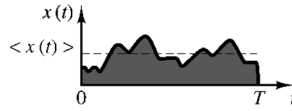



FIGURE 11.47 Averaging a signal waveform.

A circuit containing energy-storage elements is described by a differential equation. The differential equation describing the series RC circuit shown is

$$\frac{di_C}{dt} + \frac{1}{RC} i_C = \frac{dv_S}{dt}$$

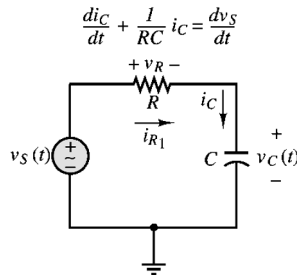



FIGURE 11.48 Circuit containing energy-storage element.

A useful measure of the voltage of an AC waveform is the rms value of the signal, $x(t)$, defined as follows:

$$x_{\text{rms}} \sqrt{\frac{1}{T}\int_0^T x^2(t)\, dt} \tag{11.47}$$

Note immediately that if $x(t)$ is a voltage, the resulting x_{rms} will also have units of volts. If you analyze Eq. (11.47), you can see that, in effect, the rms value consists of the square root of the average (or mean) of the square of the signal. Thus, the notation *rms* indicates exactly the operations performed on $x(t)$ in order to obtain its rms value.

Solution of Circuits Containing Dynamic Elements

The major difference between the analysis of the resistive circuits and circuits containing capacitors and inductors is now that the equations that result from applying Kirchhoff's laws are differential equations, as opposed to the algebraic equations obtained in solving resistive circuits. Consider, for example, the circuit of Fig. 11.48 which consists of the series connection of a voltage source, a resistor, and a capacitor. Applying KVL around the loop, we may obtain the following equation:

$$v_S(t) = v_R(t) + v_C(t) \tag{11.48}$$

Observing that $i_R = i_C$, Eq. (11.48) may be combined with the defining equation for the capacitor (Eq. 4.6.6) to obtain

$$v_S(t) = Ri_C(t) + \frac{1}{C}\int_{-\infty}^{t} i_C\, dt \tag{11.49}$$

Equation (11.49) is an integral equation, which may be converted to the more familiar form of a differential equation by differentiating both sides of the equation, and recalling that

$$\frac{d}{dt}\left(\int_{-\infty}^{t} i_C\, dt\right) = i_C(t) \tag{11.50}$$

to obtain the following differential equation:

$$\frac{di_C}{dt} + \frac{1}{RC} i_C = \frac{1}{R}\frac{dv_S}{dt} \tag{11.51}$$

where the argument (t) has been dropped for ease of notation.

Observe that in Eq. (11.51), the independent variable is the series current flowing in the circuit, and that this is not the only equation that describes the series RC circuit. If, instead of applying KVL, for example, we had applied KCL at the node connecting the resistor to the capacitor, we would have obtained the following relationship:

$$i_R = \frac{v_S - v_C}{R} = i_C = C\frac{dv_C}{dt} \tag{11.52}$$

or

$$\frac{dv_C}{dt} + \frac{1}{RC} v_C = \frac{1}{RC} v_S \tag{11.53}$$

Note the similarity between Eqs. (11.51) and (11.53). The left-hand side of both equations is identical, except for the dependent variable, while the right-hand side takes a slightly different form. The solution of either equation is sufficient, however, to determine all voltages and currents in the circuit.

We can generalize the results above by observing that any circuit containing a single energy-storage element can be described by a differential equation of the form

$$a_1\frac{dy(t)}{dt} + a_0(t) = F(t) \tag{11.54}$$

where $y(t)$ represents the capacitor voltage in the circuit of Fig. 11.48 and where the constants a_0 and a_1 consist of combinations of circuit element parameters. Equation (11.54) is a **first-order ordinary differential equation** with constant coefficients.

Consider now a circuit that contains two energy-storage elements, such as that shown in Fig. 11.49. Application of KVL results in the following equation:

$$Ri(t) + L\frac{di(t)}{dt} + \frac{1}{C}\int_{-\infty}^{t} i(t)\,dt = v_S(t) \tag{11.55}$$

Equation (11.55) is called an integro-differential equation because it contains both an integral and a derivative. This equation can be converted into a differential equation by differentiating both sides, to obtain:

$$R\frac{di(t)}{dt} + L\frac{d^2 i(t)}{dt^2} + \frac{1}{C} i(t) = \frac{dv_S(t)}{dt} \tag{11.56}$$

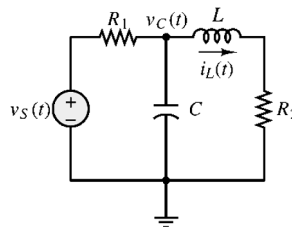



FIGURE 11.49 Second-order circuit.

or, equivalently, by observing that the current flowing in the series circuit is related to the capacitor voltage by $i(t) = Cdv_C/dt$, and that Eq. (11.55) can be rewritten as

$$RC\frac{dv_C}{dt} + LC\frac{d^2v_C(t)}{dt^2} + v_C(t) = v_S(t) \tag{11.57}$$

Note that although different variables appear in the preceding differential equations, both Eqs. (11.55) and (11.57) can be rearranged to appear in the same general form as follows:

$$a_2\frac{d^2y(t)}{dt^2} + a_1\frac{dy(t)}{dt} + a_0y(t) = F(t) \tag{11.58}$$

where the general variable $y(t)$ represents either the series current of the circuit of Fig. 11.49 or the capacitor voltage. By analogy with Eq. (11.54), we call Eq. (11.58) a **second-order ordinary differential equation** with constant coefficients. As the number of energy-storage elements in a circuit increases, one can therefore expect that higher-order differential equations will result.

Phasors and Impedance

In this section, we introduce an efficient notation to make it possible to represent sinusoidal signals as *complex numbers*, and to eliminate the need for solving differential equations.

Phasors

Let us recall that it is possible to express a generalized sinusoid as the real part of a complex vector whose **argument**, or **angle**, is given by $(\omega t + \phi)$ and whose length, or **magnitude**, is equal to the peak amplitude of the sinusoid. The **complex phasor** corresponding to the sinusoidal signal $A\cos(\omega t + \phi)$ is therefore defined to be the complex number $Ae^{j\phi}$:

$$Ae^{j\phi} = \text{complex phasor notation for } A\cos(\omega t + \phi) \tag{11.59}$$

1. Any sinusoidal signal may be mathematically represented in one of two ways: a **time-domain form**

$$v(t) = A\cos(\omega t + \phi)$$

and a **frequency-domain** (or **phasor**) **form**

$$\mathbf{V}(j\omega) = Ae^{j\phi}$$

2. A phasor is a complex number, expressed in polar form, consisting of a *magnitude* equal to the peak amplitude of the sinusoidal signal and a *phase angle* equal to the phase shift of the sinusoidal signal *referenced to a cosine signal.*
3. When using phasor notation, it is important to make a note of the specific frequency, ω, of the sinusoidal signal, since this is not explicitly apparent in the phasor expression.

Impedance

We now analyze the *i-v* relationship of the three ideal circuit elements in light of the new phasor notation. The result will be a new formulation in which resistors, capacitors, and inductors will be described in the same notation. A direct consequence of this result will be that the circuit theorems of section 11.3 will be extended to AC circuits. In the context of AC circuits, any one of the three ideal circuit elements

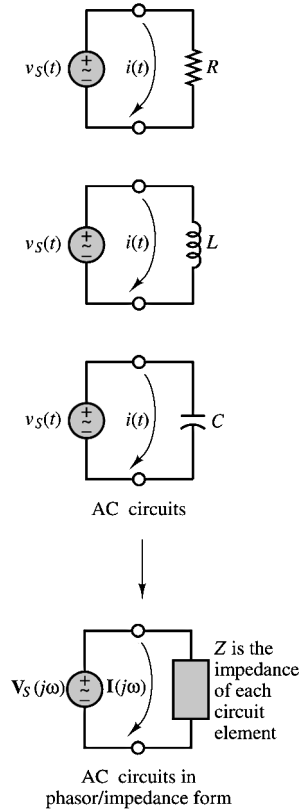



FIGURE 11.50 The impedance element.

defined so far will be described by a parameter called **impedance**, which may be viewed as a *complex resistance*. The impedance concept is equivalent to stating that capacitors and inductors act as *frequency-dependent resistors*, that is, as resistors whose resistance is a function of the frequency of the sinusoidal excitation. Figure 11.50 depicts the same circuit represented in conventional form (top) and in phasor-impedance form (bottom); the latter representation explicitly shows phasor voltages and currents and treats the circuit element as a generalized "impedance." It will presently be shown that each of the three ideal circuit elements may be represented by one such impedance element.

Let the source voltage in the circuit of Fig. 11.50 be defined by

$$v_S(t) = A\cos\omega t \quad \text{or} \quad \mathbf{V}_S(j\omega) = Ae^{j0^\circ} \tag{11.60}$$

without loss of generality. Then the current $i(t)$ is defined by the i-v relationship for each circuit element. Let us examine the frequency-dependent properties of the resistor, inductor, and capacitor, one at a time.

The *impedance* of the resistor is defined as the ratio of the phasor voltage across the resistor to the phasor current flowing through it, and the symbol Z_R is used to denote it:

$$Z_R(j\omega) = \frac{\mathbf{V}_S(j\omega)}{\mathbf{I}(j\omega)} = R \tag{11.61}$$

The impedance of the inductor is defined as follows:

$$Z_L(j\omega) = \frac{\mathbf{V}_S(j\omega)}{\mathbf{I}(j\omega)} = \omega Le^{j90^\circ} = j\omega L \tag{11.62}$$

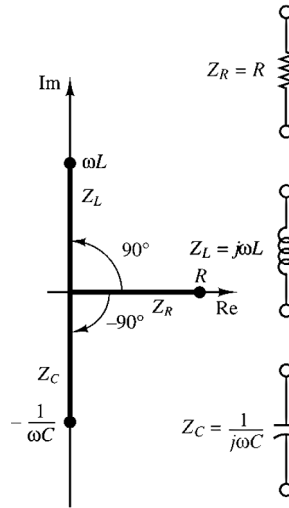



FIGURE 11.51 Impedances of R, L, and C in the complex plane.

Note that the inductor now appears to behave like a *complex frequency-dependent resistor*, and that the magnitude of this complex resistor, ωL, is proportional to the signal frequency, ω. Thus, an inductor will "impede" current flow in proportion to the sinusoidal frequency of the source signal. This means that at low signal frequencies, an inductor acts somewhat like a short circuit, while at high frequencies it tends to behave more as an open circuit. Another important point is that *the magnitude of the impedance of an inductor is always positive*, since both L and ω are positive numbers. You should verify that the units of this magnitude are also ohms.

The impedance of the ideal capacitor, $Z_C(j\omega)$, is therefore defined as follows:

$$Z_C(j\omega) = \frac{\mathbf{V}_S(j\omega)}{\mathbf{I}(j\omega)} = \frac{1}{\omega C}e^{-j90^\circ} = \frac{-j}{\omega C} = \frac{1}{j\omega C} \tag{11.63}$$

where we have used the fact that $1/j = e^{-j90^\circ} = -j$. Thus, the impedance of a capacitor is also a frequency-dependent complex quantity, with the impedance of the capacitor varying as an inverse function of frequency, and so a capacitor acts like a short circuit at high frequencies, whereas it behaves more like an open circuit at low frequencies. Another important point is that *the impedance of a capacitor is always negative*, since both C and ω are positive numbers. You should verify that the units of impedance for a capacitor are ohms. Figure 11.51 depicts $Z_C(j\omega)$ in the complex plane, alongside $Z_R(j\omega)$ and $Z_L(j\omega)$.

The impedance parameter defined in this section is extremely useful in solving AC circuit analysis problems, because it will make it possible to take advantage of most of the network theorems developed for DC circuits by replacing resistances with complex-valued impedances. In its most general form, the impedance of a circuit element is defined as the sum of a real part and an imaginary part:

$$Z(j\omega) = R(j\omega) + jX(j\omega) \tag{11.64}$$

where R is called the **AC resistance** and X is called the **reactance**. The frequency dependence of R and X has been indicated explicitly, since it is possible for a circuit to have a frequency-dependent resistance. The examples illustrate how a complex impedance containing both real and imaginary parts arises in a circuit.

Example 11.4 Capacitive Displacement Transducer

In Example 11.3, the idea of a capacitive displacement transducer was introduced when we considered a parallel-plate capacitor composed of a fixed plate and a movable plate. The capacitance of this variable capacitor was shown to be a *nonlinear* function of the position of the movable plate, x (see Fig. 11.39).

In this example, we show that under certain conditions the impedance of the capacitor varies as a *linear* function of displacement—that is, the movable-plate capacitor can serve as a linear transducer.

Recall the expression derived in Example 11.3:

$$C = \frac{8.854 \times 10^{-3} A}{x}$$

where C is the capacitance in picofarad, A is the area of the plates in square millimeter, and x is the (variable) distance in millimeter. If the capacitor is placed in an AC circuit, its impedance will be determined by the expression

$$Z_C = \frac{1}{j\omega C}$$

so that

$$Z_C = \frac{x}{8.854\, j\omega A}$$

Thus, at a fixed frequency ω, the impedance of the capacitor will vary linearly with displacement. This property may be exploited in the bridge circuit of Example 11.3, where a differential pressure transducer was shown as being made of two movable-plate capacitors, such that if the capacitance of one increased as a consequence of a pressure differential across the transducer, the capacitance of the other had to decrease by a corresponding amount (at least for small displacements). The circuit is shown again in Fig. 11.52 where two resistors have been connected in the bridge along with the variable capacitors (denoted by $C(x)$). The bridge is excited by a sinusoidal source.

Using phasor notation, we can express the output voltage as follows:

$$\mathbf{V}_{\text{out}}(j\omega) = \mathbf{V}_S(j\omega)\left(\frac{Z_{C_{bc}(x)}}{Z_{C_{db}(x)} + Z_{C_{bc}(x)}} - \frac{R_2}{R_1 + R_2}\right)$$

If the nominal capacitance of each movable-plate capacitor with the diaphragm in the center position is given by

$$C = \frac{\varepsilon A}{d}$$

where d is the nominal (undisplaced) separation between the diaphragm and the fixed surfaces of the capacitors (in mm), the capacitors will see a change in capacitance given by

$$C_{db} = \frac{\varepsilon A}{d - x} \quad \text{and} \quad C_{bc} = \frac{\varepsilon A}{d + x}$$

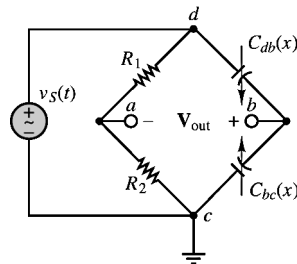



FIGURE 11.52 Bridge circuit for capacitive displacement transducer.

when a pressure differential exists across the transducer, so that the impedances of the variable capacitors change according to the displacement

$$Z_{C_{db}} = \frac{d - x}{8.854\, j\omega A} \quad \text{and} \quad Z_{C_{bc}} = \frac{d + x}{8.854\, j\omega A}$$

and we obtain the following expression for the phasor output voltage, if we choose $R_1 = R_2$.

$$\begin{aligned}
\mathbf{V}_{\text{out}}(j\omega) &= \mathbf{V}_S(j\omega)\left(\frac{\dfrac{d+x}{8.854\, j\omega A}}{\dfrac{d-x}{8.854\, j\omega A} + \dfrac{d+x}{8.854\, j\omega A}} - \frac{R_2}{R_1 + R_2}\right) \\
&= \mathbf{V}_S(j\omega)\left(\frac{1}{2} + \frac{x}{2d} - \frac{R_2}{R_1 + R_2}\right) \\
&= \mathbf{V}_S(j\omega)\frac{x}{2d}
\end{aligned}$$

Thus, the output voltage will vary as a scaled version of the input voltage in proportion to the displacement.

References

Irwin, J.D., 1989. *Basic Engineering Circuit Analysis,* 3rd ed., Macmillan, New York.

Nilsson, J.W., 1989. *Electric Circuits,* 3rd ed., Addison-Wesley, Reading, MA.

Rizzoni, G., 2000. *Principles and Applications of Electrical Engineering,* 3rd ed., McGraw-Hill, Burr Ridge, IL.

Smith, R.J. and Dorf, R.C., 1992. *Circuits, Devices and Systems,* 5th ed., John Wiley & Sons, New York.

1993. *The Electrical Engineering Handbook,* CRC Press, Boca Raton, FL.

Budak, A., *Passive and Active Network Analysis and Synthesis,* Houghton Mifflin, Boston.

Van Valkenburg, M.E., 1982, *Analog Filter Design,* Holt, Rinehart & Winston, New York.

12
Engineering Thermodynamics

Michael J. Moran
The Ohio State University

Although various aspects of what is now known as thermodynamics have been of interest since antiquity, formal study began only in the early nineteenth century through consideration of the motive power of heat: the capacity of hot bodies to produce work. Today the scope is larger, dealing generally with energy and entropy, and with relationships among the properties of matter. Moreover, in the past 25 years engineering thermodynamics has undergone a revolution, both in terms of the presentation of fundamentals and in the manner that it is applied. In particular, the second law of thermodynamics has emerged as an effective tool for engineering analysis and design.

12.1 Fundamentals

Classical thermodynamics is concerned primarily with the macrostructure of matter. It addresses the gross characteristics of large aggregations of molecules and not the behavior of individual molecules. The microstructure of matter is studied in kinetic theory and statistical mechanics (including quantum thermodynamics). In this chapter, the classical approach to thermodynamics is featured.

Basic Concepts and Definitions

Thermodynamics is both a branch of physics and an engineering science. The scientist is normally interested in gaining a fundamental understanding of the physical and chemical behavior of fixed, quiescent quantities of matter and uses the principles of thermodynamics to relate the properties of matter. Engineers are generally interested in studying systems and how they interact with their surroundings. To facilitate this, engineers have extended the subject of thermodynamics to the study of systems through which matter flows.

System

In a thermodynamic analysis, the *system* is the subject of the investigation. Normally the system is a specified quantity of matter and/or a region that can be separated from everything else by a well-defined surface. The defining surface is known as the *control surface* or *system boundary*. The control surface may be movable or fixed. Everything external to the system is the *surroundings*. A system of fixed mass is

referred to as a *control mass* or *closed system*. When there is flow of mass through the control surface, the system is called a *control volume* or *open system*. An *isolated* system is a closed system that does not interact in any way with its surroundings.

State, Property

The condition of a system at any instant of time is called its *state*. The state at a given instant of time is described by the properties of the system. A *property* is any quantity whose numerical value depends on the state, but not the history of the system. The value of a property is determined in principle by some type of physical operation or test.

Extensive properties depend on the size or extent of the system. Volume, mass, energy, entropy, and exergy are examples of extensive properties. An extensive property is additive in the sense that its value for the whole system equals the sum of the values for its parts. *Intensive* properties are independent of the size or extent of the system. Pressure and temperature are examples of intensive properties.

Process, Cycle

Two states are identical if, and only if, the properties of the two states are identical. When any property of a system changes in value there is a change in state, and the system is said to undergo a *process*. When a system in a given initial state goes through a sequence of processes and finally returns to its initial state, it is said to have undergone a *thermodynamic cycle*.

Phase and Pure Substance

The term *phase* refers to a quantity of matter that is homogeneous throughout in both chemical composition and physical structure. Homogeneity in physical structure means that the matter is all *solid*, or all *liquid*, or all *vapor* (or equivalently all *gas*). A system can contain one or more phases. For example, a system of liquid water and water vapor (steam) contains two phases. A *pure substance* is one that is uniform and invariable in chemical composition. A pure substance can exist in more than one phase, but its chemical composition must be the same in each phase. For example, if liquid water and water vapor form a system with two phases, the system can be regarded as a pure substance because each phase has the same composition. The nature of phases that coexist in equilibrium is addressed by the *phase rule* (for discussion see Moran and Shapiro, 2000).

Equilibrium

Equilibrium means a condition of balance. In thermodynamics the concept includes not only a balance of forces, but also a balance of other influences. Each kind of influence refers to a particular aspect of thermodynamic (complete) equilibrium. *Thermal* equilibrium refers to an equality of temperature, *mechanical* equilibrium to an equality of pressure, and *phase* equilibrium to an equality of chemical potentials (for discussion see Moran and Shapiro, 2000). *Chemical* equilibrium is also established in terms of chemical potentials. For complete equilibrium the several types of equilibrium must exist individually.

Temperature

A scale of temperature independent of the thermometric substance is called a *thermodynamic temperature scale*. The Kelvin scale, a thermodynamic scale, can be elicited from the second law of thermodynamics. The definition of temperature following from the second law is valid over all temperature ranges and provides an essential connection between the several *empirical* measures of temperature. In particular, temperatures evaluated using a *constant-volume gas thermometer* are identical to those of the Kelvin scale over the range of temperatures where gas thermometry can be used. On the Kelvin scale the unit is the kelvin (K).

The Celsius temperature scale (also called the centigrade scale) uses the degree Celsius (∞C), which has the same magnitude as the kelvin. Thus, temperature differences are identical on both scales. However, the zero point on the Celsius scale is shifted to 273.15 K, the *triple point* of water (Fig. 12.1b),

as shown by the following relationship between the Celsius temperature and the Kelvin temperature:

$$T(^\circ\mathrm{C}) = T(\mathrm{K}) - 273.15 \tag{12.1}$$

Two other temperature scales are commonly used in engineering in the U.S. By definition, the *Rankine scale*, the unit of which is the degree rankine (°R), is proportional to the Kelvin temperature according to

$$T(^\circ\mathrm{R}) = 1.8T(\mathrm{K}) \tag{12.2}$$

The Rankine scale is also an absolute thermodynamic scale with an absolute zero that coincides with the absolute zero of the Kelvin scale. In thermodynamic relationships, temperature is always in terms of the Kelvin or Rankine scale unless specifically stated otherwise.

A degree of the same size as that on the Rankine scale is used in the *Fahrenheit scale*, but the zero point is shifted according to the relation

$$T(^\circ\mathrm{F}) = T(^\circ\mathrm{R}) - 459.67 \tag{12.3}$$

Substituting Eqs. (12.1) and (12.2) into Eq. (12.3) gives

$$T(^\circ\mathrm{F}) = 1.8T(^\circ\mathrm{C}) + 32 \tag{12.4}$$

This equation shows that the Fahrenheit temperature of the *ice point* (0°C) is 32°F and of the *steam point* (100°C) is 212°F. The 100 Celsius or Kelvin degrees between the ice point and steam point corresponds to 180 Fahrenheit or Rankine degrees.

To provide a standard for temperature measurement taking into account both theoretical and practical considerations, the International Temperature Scale of 1990 (ITS-90) is defined in such a way that the temperature measured on it conforms with the thermodynamic temperature, the unit of which is the kelvin, to within the limits of accuracy of measurement obtainable in 1990. Further discussion of ITS-90 is provided by Preston-Thomas (1990).

Irreversibilities

A process is said to be *reversible* if it is possible for its effects to be eradicated in the sense that there is some way by which both the system and its surroundings can be exactly restored to their respective initial states. A process is *irreversible* if both the system and surroundings cannot be restored to their initial states. There are many effects whose presence during a process renders it irreversible. These include, but are not limited to, the following: heat transfer through a finite temperature difference; unrestrained expansion of a gas or liquid to a lower pressure; spontaneous chemical reaction; mixing of matter at different compositions or states; friction (sliding friction as well as friction in the flow of fluids); electric current flow through a resistance; magnetization or polarization with hysteresis; and inelastic deformation. The term *irreversibility* is used to identify effects such as these.

Irreversibilities can be divided into two classes, *internal* and *external*. Internal irreversibilities are those that occur within the system, while external irreversibilities are those that occur within the surroundings, normally the immediate surroundings. As this division depends on the location of the boundary there is some arbitrariness in the classification (by locating the boundary to take in the immediate surroundings, all irreversibilities are internal). Nonetheless, valuable insights can result when this distinction between irreversibilities is made. When internal irreversibilities are absent during a process, the process is said to be *internally reversible*. At every intermediate state of an internally reversible process of a closed system, all intensive properties are uniform throughout each phase present: the temperature, pressure, specific volume, and other intensive properties do not vary with position.

Laws of Thermodynamics

The first steps in a thermodynamic analysis are definition of the system and identification of the relevant interactions with the surroundings. Attention then turns to the pertinent physical laws and relationships that allow the behavior of the system to be described in terms of an engineering model, which is a simplified representation of system behavior that is sufficiently faithful for the purpose of the analysis, even if features exhibited by the actual system are ignored.

Thermodynamic analyses of control volumes and closed systems typically use, directly or indirectly, one or more of three basic laws. The laws, which are independent of the particular substance or substances under consideration, are

- the conservation of mass principle,
- the conservation of energy principle,
- the second law of thermodynamics.

The second law may be expressed in terms of entropy or exergy.

The laws of thermodynamics must be supplemented by appropriate thermodynamic property data. For some applications a momentum equation expressing Newton's second law of motion also is required. Data for transport properties, heat transfer coefficients, and friction factors often are needed for a comprehensive engineering analysis. Principles of engineering economics and pertinent economic data also can play prominent roles.

12.2 Extensive Property Balances

The laws of thermodynamics can be expressed in terms of *extensive property balances* for mass, energy, entropy, and exergy. Engineering applications are generally analyzed on a control volume basis. Accordingly, the control volume formulations of the mass energy, entropy, and exergy balances are featured here. They are provided in the form of overall balances assuming one-dimensional flow. Equations of change for mass, energy, and entropy in the form of differential equations are also available in the literature (Bird et al., 1960).

Mass Balance

For applications in which inward and outward flows occur, each through one or more ports, the extensive property balance expressing the conservation of mass principle takes the form

$$\frac{dm}{dt} = \sum_i \dot{m}_i - \sum_e \dot{m}_e \tag{12.5}$$

where dm/dt represents the time rate of change of mass contained within the control volume, $\dot{m}_i$ denotes the mass flow rate at an inlet port, and $\dot{m}_e$ denotes the mass flow rate at an exit port.

The volumetric flow rate through a portion of the control surface with area dA is the product of the velocity component normal to the area, v_n, times the area: $\mathrm{v}_n dA$. The mass flow rate through dA is $\rho(\mathrm{v}_n dA)$, where ρ denotes density. The mass rate of flow through a port of area A is then found by integration over the area

$$\dot{m} = \int_A \rho \mathrm{v}_n \, dA$$

For one-dimensional flow the intensive properties are uniform with position over area A, and the last equation becomes

$$\dot{m} = \rho \mathrm{v} A = \frac{\mathrm{v}A}{v} \tag{12.6}$$

where v denotes the specific volume (the reciprocal of density) and the subscript n has been dropped from velocity for simplicity.

Energy Balance

Energy is a fundamental concept of thermodynamics and one of the most significant aspects of engineering analysis. Energy can be stored within systems in various macroscopic forms: kinetic energy, gravitational potential energy, and internal energy. Energy also can be transformed from one form to another and transferred between systems. Energy can be transferred by work, by heat transfer, and by flowing matter. The total amount of energy is conserved in all transformations and transfers. The extensive property balance expressing the conservation of energy principle takes the form

$$\frac{d(U + \mathrm{KE} + \mathrm{PE})}{dt} = \dot{Q} - \dot{W} + \sum_i \dot{m}_i\left(h_i + \frac{\mathrm{v}_i^2}{2} + gz_i\right) - \sum_e \dot{m}_e\left(h_e + \frac{\mathrm{v}_e^2}{2} + gz_e\right) \tag{12.7a}$$

where U, KE, and PE denote, respectively, the internal energy, kinetic energy, and gravitational potential energy of the overall control volume.

The right side of Eq. (12.7a) accounts for transfers of energy across the boundary of the control volume. Energy can enter and exit control volumes by work. Because work is done on or by a control volume when matter flows across the boundary, it is convenient to separate the work rate (or power) into two contributions. One contribution is the work rate associated with the force of the fluid pressure as mass is introduced at the inlet and removed at the exit. Commonly referred to as *flow work*, this contribution is accounted for by $\dot{m}_i(p_i v_i)$ and $\dot{m}_e(p_e v_e)$, respectively, where p denotes pressure and v denotes specific volume. The other contribution, denoted by $\dot{W}$ in Eq. (12.7a), includes all other work effects, such as those associated with rotating shafts, displacement of the boundary, and electrical effects. $\dot{W}$ is considered *positive* for energy transfer *from* the control volume.

Energy also can enter and exit control volumes with flowing streams of matter. On a one-dimensional flow basis, the rate at which energy enters with matter at inlet i is $\dot{m}_i(u_i + \mathrm{v}_i^2/2 + gz_i)$, where the three terms in parentheses account, respectively, for the specific internal energy, specific kinetic energy, and specific gravitational potential energy of the substance flowing through port i. In writing Eq. (12.7a) the sum of the specific internal energy and specific flow work at each inlet and exit is expressed in terms of the specific enthalpy $h(=u + pv)$. Finally, $\dot{Q}$ accounts for the rate of energy transfer by heat and is considered *positive* for energy transfer *to* the control volume.

By dropping the terms of Eq. (12.7a) involving mass flow rates an energy rate balance for closed systems is obtained. In principle the closed system energy rate balance can be integrated for a process between two states to give the closed system energy balance:

$$\underset{\text{(closed systems)}}{(U_2 - U_1) + (\mathrm{KE}_2 - \mathrm{KE}_1) + (\mathrm{PE}_2 - \mathrm{PE}_1) = Q - W} \tag{12.7b}$$

where 1 and 2 denote the end states. Q and W denote the *amounts* of energy transferred by heat and work during the process, respectively.

Entropy Balance

Contemporary applications of engineering thermodynamics express the second law, alternatively, as an entropy balance or an exergy balance. The entropy balance is considered here.

Like mass and energy, entropy can be stored within systems and transferred across system boundaries. However, unlike mass and energy, entropy is not conserved, but generated (or produced) by *irreversibilities*

within systems. A control volume form of the extensive property balance for entropy is

$$\frac{dS}{dt} = \underbrace{\sum_j \frac{\dot{Q}_j}{T_j} + \sum_i \dot{m}_i s_i - \sum_e \dot{m}_e s_e}_{\text{rates of entropy transfer}} + \underbrace{\dot{S}_{\text{gen}}}_{\text{rate of entropy generation}} \quad (12.8)$$

where dS/dt represents the time rate of change of entropy within the control volume. The terms $\dot{m}_i s_i$ and $\dot{m}_e s_e$ account, respectively, for rates of entropy transfer into and out of the control volume accompanying mass flow. $\dot{Q}_j$ represents the time rate of heat transfer at the location on the boundary where the instantaneous temperature is T_j, and $\dot{Q}_j/T_j$ accounts for the accompanying rate of entropy transfer. $\dot{S}_{\text{gen}}$ denotes the time rate of entropy generation due to irreversibilities within the control volume. An entropy rate balance for closed systems is obtained by dropping the terms of Eq. (12.8) involving mass flow rates.

When applying the entropy balance in any of its forms, the objective is often to evaluate the entropy generation term. However, the value of the entropy generation for a given process of a system usually does not have much significance by itself. The significance normally is determined through comparison: the entropy generation within a given component would be compared with the entropy generation values of the other components included in an overall system formed by these components. This allows the principal contributors to the irreversibility of the overall system to be pinpointed.

Control Volumes at Steady State

Engineering systems are often idealized as being at *steady state*, meaning that all properties are unchanging in time. For a control volume at steady state, the identity of the matter within the control volume changes continuously, but the total amount of mass remains constant. At steady state, the mass rate balance Eq. (12.5) reduces to

$$\sum_i \dot{m}_i = \sum_e \dot{m}_e \quad (12.9a)$$

At steady state, the energy rate balance Eq. (12.7a) becomes

$$0 = \dot{Q} - \dot{W} + \sum_i \dot{m}_i \left(h_i + \frac{V_i^2}{2} + gz_i \right) - \sum_e \dot{m}_e \left(h_e + \frac{V_e^2}{2} + gz_e \right) \quad (12.9b)$$

At steady state, the entropy rate balance Eq. (12.8) reads

$$0 = \sum_j \frac{\dot{Q}_j}{T_j} + \sum_i \dot{m}_i s_i - \sum_e \dot{m}_e s_e + \dot{S}_{\text{gen}} \quad (12.9c)$$

Mass and energy are conserved quantities, but entropy is not generally conserved. Equation (12.9a) indicates that the total rate of mass flow into the control volume equals the total rate of mass flow out of the control volume. Similarly, Eq. (12.9b) states that the total rate of energy transfer into the control volume equals the total rate of energy transfer out of the control volume. However, Eq. (12.9c) shows that the rate at which entropy is transferred out exceeds the rate at which entropy enters, the difference being the rate of entropy generation within the control volume owing to irreversibilities.

Many applications involve control volumes having a single inlet and a single exit. For such cases the mass rate balance, Eq. (12.9a), reduces to $\dot{m}_i = \dot{m}_e$. Denoting the common mass flow rate by $\dot{m}$,

Eqs. (12.9b) and (12.9c) give, respectively,

$$0 = \dot{Q} - \dot{W} + \dot{m}\left[(h_i - h_e) + \left(\frac{V_i^2 - V_e^2}{2}\right) + g(z_i - z_e)\right] \tag{12.10a}$$

$$0 = \frac{\dot{Q}}{T_b} + \dot{m}(s_i - s_e) + \dot{S}_{gen} \tag{12.11a}$$

where for simplicity T_b denotes the temperature, or a suitable average temperature, on the boundary where heat transfer occurs.

When energy and entropy rate balances are applied to particular cases of interest, additional simplifications are usually made. The heat transfer term $\dot{Q}$ is dropped when it is insignificant relative to other energy transfers across the boundary. This may be the result of one or more of the following: (1) the outer surface of the control volume is insulated; (2) the outer surface area is too small for there to be effective heat transfer; (3) the temperature difference between the control volume and its surroundings is small enough that the heat transfer can be ignored; (4) the gas or liquid passes through the control volume so quickly that there is not enough time for significant heat transfer to occur. The work term $\dot{W}$ drops out of the energy rate balance when there are no rotating shafts, displacements of the boundary, electrical effects, or other work mechanisms associated with the control volume being considered. The effects of kinetic and potential energy are frequently negligible relative to other terms of the energy rate balance.

The special forms of Eqs. (12.10a) and (12.11a) listed in Table 12.1 are obtained as follows: When there is no heat transfer, Eq. (12.11a) gives

$$s_e - s_i = \frac{\dot{S}_{gen}}{\dot{m}} \geq 0 \tag{12.11b}$$

(no heat transfer)

Accordingly, when irreversibilities are present within the control volume, the specific entropy increases as mass flows from inlet to outlet. In the *ideal* case in which no internal irreversibilities are present, mass passes through the control volume with no change in its entropy—that is, *isentropically.*

For no heat transfer, Eq. (12.10a) gives

$$\dot{W} = \dot{m}\left[(h_i - h_e) + \left(\frac{V_i^2 - V_e^2}{2}\right) + g(z_i - z_e)\right] \tag{12.10b}$$

(no heat transfer)

A special form that is applicable, at least approximately, to compressors, pumps, and turbines results from dropping the kinetic and potential energy terms of Eq. (12.10b), leaving

$$\dot{W} = \dot{m}(h_i - h_e) \tag{12.10c}$$

(compressors, pumps, and turbines)

In *throttling devices* a significant reduction in pressure is achieved by introducing a restriction into a line through which a gas or liquid flows. For such devices $\dot{W} = 0$ and Eq. (12.10c) reduces further to read

$$h_i \cong h_e \tag{12.10d}$$

(throttling process)

That is, upstream and downstream of the throttling device, the specific enthalpies are equal.

TABLE 12.1 Energy and Entropy Balances for One-Inlet, One-Outlet Control Volumes at Steady State and No Heat Transfer

Energy balance

$$\dot{W} = \dot{m}\left[(h_i - h_e) + \left(\frac{v_i^2 - v_e^2}{2}\right) + g(z_i - z_e)\right] \quad (12.10b)$$

Compressors, pumps, and turbines[a]

$$\dot{W} = \dot{m}(h_i - h_e) \quad (12.10c)$$

Throttling

$$h_e \cong h_i \quad (12.10d)$$

Nozzles, diffusers[b]

$$v_e = \sqrt{v_i^2 + 2(h_i - h_e)} \quad (12.10e)$$

Entropy balance

$$s_e - s_i = \frac{\dot{S}_{gen}}{\dot{m}} \geq 0 \quad (12.11b)$$

[a] For an ideal gas with constant c_p, Eq. (1′) of Table 12.4 allows Eq. (12.10c) to be written as

$$\dot{W} = \dot{m}c_p(T_i - T_e) \quad (12.10c')$$

The power developed in an *isentropic process* is obtained with Eq. (5′) of Table 12.4 as

$$\dot{W} = \dot{m}c_p T_i[1 - (p_e/p_i)^{(k-1)/k}] \quad (s = c) \quad (12.10c'')$$

where $c_p = kR/(k-1)$.

[b] For an ideal gas with constant c_p, Eq. (1′) of Table 12.4 allows Eq. (12.10e) to be written as

$$v_e = \sqrt{v_i^2 + 2c_p(T_i - T_e)} \quad (12.10e')$$

The exit velocity for an *isentropic process* is obtained with Eq. (5′) of Table 12.4 as

$$v_e = \sqrt{v_i^2 + 2c_p T_i[1 - (p_e/p_i)^{(k-1)/k}]} \quad (s = c) \quad (12.10e'')$$

where $c_p = kR/(k - 1)$.

A *nozzle* is a flow passage of varying cross-sectional area in which the velocity of a gas or liquid increas es in the direction of flow. In a *diffuser*, the gas or liquid decelerates in the direction of flow. For such devices, $\dot{W} = 0$. The heat transfer and potential energy change are generally negligible. Then Eq. (12.10b) reduces to

$$0 = h_i - h_e + \frac{v_i^2 - v_e^2}{2}$$

Solving for the exit velocity

$$v_e = \sqrt{v_i^2 + 2(h_i - h_e)} \quad (12.10e)$$

(nozzle, diffuser)

The steady-state forms of the mass, energy, and entropy rate balances can be applied to control volumes with multiple inlets and/or exits, for example, cases involving heat-recovery steam generators, feedwater heaters, and counterflow and crossflow heat exchangers. Transient (or unsteady) analyses can be conducted with Eqs. (12.5), (12.7a), and (12.8). Illustrations of all such applications are provided by Moran and Shapiro (2000).

Exergy Balance

Exergy provides an alternative to entropy for applying the second law. When exergy concepts are combined with principles of engineering economy, the result is known as *thermoeconomics*. Thermoeconomics allows the real cost sources to be identified: capital investment costs, operating and maintenance costs, and the costs associated with the destruction and loss of exergy. Optimization of systems can be achieved by a careful consideration of such cost sources. From this perspective thermoeconomics is *exergy-aided cost minimization*. Discussions of exergy analysis and thermoeconomics are provided by Moran (1989), Bejan et al. (1996), Moran and Tsatsaronis (2000), and Moran and Shapiro (2000). In this section salient aspects are presented.

Defining Exergy

An opportunity for doing work exists whenever two systems at different states are placed in communication because, in principle, work can be developed as the two are allowed to come into equilibrium. When one of the two systems is a suitably idealized system called an *environment* and the other is some system of interest, *exergy* is the maximum theoretical useful work (shaft work or electrical work) obtainable as the system of interest and environment interact to equilibrium, heat transfer occurring with the environment only. (Alternatively, exergy is the minimum theoretical useful work required to form a quantity of matter from substances present in the environment and bring the matter to a specified state.) Exergy is a measure of the *departure* of the state of the system from that of the environment, and is therefore an attribute of the system and environment together. Once the environment is specified, however, a value can be assigned to exergy in terms of property values for the system only, so exergy can be regarded as an extensive property of the system. Exergy can be destroyed and, like entropy, generally is not conserved.

Models with various levels of specificity are employed for describing the environment used to evaluate exergy. Models of the environment typically refer to some portion of a system's surroundings, the intensive properties of each phase of which are uniform and do not change significantly as a result of any process under consideration. The environment is regarded as composed of common substances existing in abundance within the Earth's atmosphere, oceans, and crust. The substances are in their stable forms as they exist naturally, and there is no possibility of developing work from interactions—physical or chemical—between parts of the environment. Although the intensive properties of the environment are assumed to be unchanging, the extensive properties can change as a result of interactions with other systems. Kinetic and potential energies are evaluated relative to coordinates in the environment, all parts of which are considered to be at rest with respect to one another. For computational ease, the temperature T_0 and pressure p_0 of the environment are often taken as typical ambient values, such as 1 atm and 25°C (77°F). However, these properties may be specified differently depending on the application.

When a system is in equilibrium with the environment, the state of the system is called the *dead state*. At the dead state, the conditions of mechanical, thermal, and chemical equilibrium between the system and the environment are satisfied: the pressure, temperature, and chemical potentials of the system equal those of the environment, respectively. In addition, the system has no motion or elevation relative to coordinates in the environment. Under these conditions, there is no possibility of a spontaneous change within the system or the environment, nor can there be an interaction between them. The value of exergy is zero. Another type of equilibrium between the system and environment can be identified. This is a restricted form of equilibrium where only the conditions of mechanical and thermal equilibrium must be satisfied. This state of the system is called the *restricted dead state*. At the restricted dead state, the fixed quantity of matter under consideration is imagined to be sealed in an envelope impervious to mass flow, at zero velocity and elevation relative to coordinates in the environment, and at the temperature T_0 and pressure p_0.

Exergy Transfer and Exergy Destruction

Exergy can be transferred by three means: exergy transfer associated with work, exergy transfer associated with heat transfer, and exergy transfer associated with the matter entering and exiting a control volume. All such exergy transfers are evaluated relative to the environment used to define exergy. Exergy also is

destroyed by irreversibilities within the system or control volume. Exergy balances can be written in various forms, depending on whether a closed system or control volume is under consideration and whether steady-state or transient operation is of interest. Owing to its importance for a wide range of applications, an exergy rate balance for control volumes at steady state is presented alternatively as Eqs. (12.12a) and (12.12b).

$$0 = \underbrace{\sum_j \dot{E}_{q,j} - \dot{W} + \sum_i \dot{E}_i - \sum_e \dot{E}_e}_{\text{rates of exergy transfer}} - \underbrace{\dot{E}_D}_{\text{rate of exergy destruction}} \tag{12.12a}$$

$$0 = \sum_j \left(1 - \frac{T_0}{T_j}\right)\dot{Q}_j - \dot{W} + \sum_i \dot{m}_i e_i - \sum_e \dot{m}_e e_e - \dot{E}_D \tag{12.12b}$$

$\dot{W}$ has the same significance as in Eq. (12.7a): the work rate excluding the flow work. $\dot{Q}_j$ is the time rate of heat transfer at the location on the boundary of the control volume where the instantaneous temperature is T_j. The associated rate of exergy transfer is

$$\dot{E}_{q,j} = \left(1 - \frac{T_0}{T_j}\right)\dot{Q}_j \tag{12.13}$$

As for other control volume rate balances, the subscripts i and e denote inlets and exits, respectively. The exergy transfer rates at control volume inlets and exits are denoted, respectively, as $\dot{E}_i = \dot{m}_i e_i$ and $\dot{E}_e = \dot{m}_e e_e$. Finally, $\dot{E}_D$ accounts for the time rate of exergy destruction due to irreversibilities within the control volume. The exergy destruction rate is related to the entropy generation rate by

$$\dot{E}_D = T_0 \dot{S}_{\text{gen}} \tag{12.14}$$

The specific exergy transfer terms e_i and e_e are expressible in terms of four components: physical exergy e^{PH}, kinetic exergy e^{KN}, potential exergy e^{PT}, and chemical exergy e^{CH}:

$$e = e^{\text{PH}} + e^{\text{KN}} + e^{\text{PT}} + e^{\text{CH}} \tag{12.15a}$$

The first three components are evaluated as follows:

$$e^{\text{PH}} = (h - h_0) - T_0(s - s_0) \tag{12.15b}$$

$$e^{\text{KN}} = \frac{1}{2}\text{v}^2 \tag{12.15c}$$

$$e^{\text{PT}} = gz \tag{12.15d}$$

In Eq. (12.15b), h_0 and s_0 denote, respectively, the specific enthalpy and specific entropy at the restricted dead state. In Eqs. (12.15c) and (12.15d), v and z denote velocity and elevation relative to coordinates in the environment, respectively.

To evaluate the chemical exergy (the exergy component associated with the departure of the chemical composition of a system from that of the environment), alternative models of the environment can be employed depending on the application; see for example Moran (1989) and Kotas (1995). Exergy analysis is facilitated, however, by employing a *standard environment* and a corresponding table of *standard*

chemical exergies. Standard chemical exergies are based on standard values of the environmental temperature T_0 and pressure p_0 — for example, 298.15 K (25°C) and 1 atm, respectively. Standard environments also include a set of reference substances with standard concentrations reflecting as closely as possible the chemical makeup of the natural environment. Standard chemical exergy data is provided by Szargut et al. (1988), Bejan et al. (1996), and Moran and Shapiro (2000).

Guidelines for Improving Thermodynamic Effectiveness

To improve thermodynamic effectiveness it is necessary to deal directly with inefficiencies related to exergy destruction and exergy loss. The primary contributors to exergy destruction are chemical reaction, heat transfer, mixing, and friction, including unrestrained expansions of gases and liquids. To deal with them effectively, the principal sources of inefficiency not only should be understood qualitatively, but also determined quantitatively, at least approximately. Design changes to improve effectiveness must be done judiciously, however, for the cost associated with different sources of inefficiency can be different. For example, the unit cost of the electrical or mechanical power required to provide for the exergy destroyed owing to a pressure drop is generally higher than the unit cost of the fuel required for the exergy destruction caused by combustion or heat transfer.

Chemical reaction is a significant source of thermodynamic inefficiency. Accordingly, it is generally good practice to minimize the use of combustion. In many applications the use of combustion equipment such as boilers is unavoidable, however. In these cases a significant reduction in the combustion irreversibility by conventional means simply cannot be expected, for the major part of the exergy destruction introduced by combustion is an inevitable consequence of incorporating such equipment. Still, the exergy destruction in practical combustion systems can be reduced by minimizing the use of excess air and by preheating the reactants. In most cases only a small part of the exergy destruction in a combustion chamber can be avoided by these means. Consequently, after considering such options for reducing the exergy destruction related to combustion, efforts to improve thermodynamic performance should focus on components of the overall system that are more amenable to betterment by cost-effective measures. In other words, some exergy destructions and energy losses can be avoided, others cannot. Efforts should be centered on those that can be avoided.

Nonidealities associated with heat transfer also typically contribute heavily to inefficiency. Accordingly, unnecessary or cost-ineffective heat transfer must be avoided. Additional guidelines follow:

- The higher the temperature T at which a heat transfer occurs in cases where $T > T_0$, where T_0 denotes the temperature of the environment, the more valuable the heat transfer and, consequently, the greater the need to avoid heat transfer to the ambient, to cooling water, or to a refrigerated stream. Heat transfer across T_0 should be avoided.
- The lower the temperature T at which a heat transfer occurs in cases where $T < T_0$, the more valuable the heat transfer and, consequently, the greater the need to avoid direct heat transfer with the ambient or a heated stream.
- Since exergy destruction associated with heat transfer between streams varies inversely with the temperature level, the lower the temperature level, the greater the need to minimize the stream-to-stream temperature difference.

Although irreversibilities related to friction, unrestrained expansion, and mixing are often less significant than combustion and heat transfer, they should not be overlooked, and the following guidelines apply:

- Relatively more attention should be paid to the design of the lower temperature stages of turbines and compressors (the last stages of turbines and the first stages of compressors) than to the remaining stages of these devices. For turbines, compressors, and motors, consider the most thermodynamically efficient options.
- Minimize the use of throttling; check whether power recovery expanders are a cost-effective alternative for pressure reduction.

TABLE 12.2 Symbols and Definitions for Selected Properties

Property	Symbol	Definition	Property	Symbol	Definition
Pressure	p		Specific heat, constant volume	c_v	$(\partial u/\partial T)_v$
Temperature	T		Specific heat, constant pressure	c_p	$(\partial h/\partial T)_p$
Specific volume	v		Volume expansivity	β	$\frac{1}{v}(\partial v/\partial T)_p$
Specific internal energy	u		Isothermal compressivity	κ	$-\frac{1}{v}(\partial v/\partial p)_T$
Specific entropy	s		Isentropic compressibility	α	$-\frac{1}{v}(\partial v/\partial p)_s$
Specific enthalpy	h	$u + pv$	Isothermal bulk modulus	B	$-v(\partial p/\partial v)_T$
Specific Helmholtz function	ψ	$u - Ts$	Isentropic bulk modulus	B_s	$-v(\partial p/\partial v)_s$
Specific Gibbs function	g	$h - Ts$	Joule–Thomson coefficient	μ_J	$(\partial T/\partial p)_h$
Compressibility factor	Z	pv/RT	Joule coefficient	η	$(\partial T/\partial v)_u$
Specific heat ratio	k	c_p/c_v	Velocity of sound	c	$\sqrt{-v^2(\partial p/\partial v)_s}$

- Avoid processes using excessively large thermodynamic driving forces (differences in temperature, pressure, and chemical composition). In particular, minimize the mixing of streams differing significantly in temperature, pressure, or chemical composition.
- The greater the mass flow rate the greater the need to use the exergy of the stream effectively.

Discussion of means for improving thermodynamic effectiveness also is provided by Bejan et al. (1996) and Moran and Tsatsaronis (2000).

12.3 Property Relations and Data

Engineering thermodynamics uses a wide assortment of thermodynamic properties and relations among these properties. Table 12.2 lists several commonly encountered properties. Pressure, temperature, and specific volume can be found experimentally. Specific internal energy, entropy, and enthalpy are among those properties that are not so readily obtained in the laboratory. Values for such properties are calculated using experimental data of properties that are more amenable to measurement, together with appropriate property relations derived using the principles of thermodynamics.

Property data are provided in the publications of the National Institute of Standards and Technology (formerly the U.S. Bureau of Standards), of professional groups such as the American Society of Mechanical Engineers (ASME), the American Society of Heating, Refrigerating, and Air Conditioning Engineers (ASHRAE), and the American Chemical Society, and of corporate entities such as Dupont and Dow Chemical. Handbooks and property reference volumes such as included in the list of references for this chapter are readily accessed sources of data. Property data also are retrievable from various commercial online data bases. Computer software increasingly is available for this purpose as well.

P-v-T Surface

Considerable pressure, specific volume, and temperature data have been accumulated for industrially important gases and liquids. These data can be represented in the form $p = f(v, T)$, called an *equation of state*. Equations of state can be expressed in graphical, tabular, and analytical forms. Figure 12.1(a) shows the p-v-T relationship for water. Figure 12.1(b) shows the projection of the p-v-T *surface* onto the pressure-temperature plane, called the *phase diagram*. The projection onto the p-v plane is shown in Fig. 12.1(c).

Figure 12.1(a) has three regions labeled solid, liquid, and vapor where the substance exists only in a single phase. Between the single phase regions lie *two-phase* regions, where two phases coexist in equilibrium. The lines separating the single-phase regions from the two-phase regions are *saturation lines*. Any state represented by a point on a saturation line is a *saturation state*. The line separating the liquid

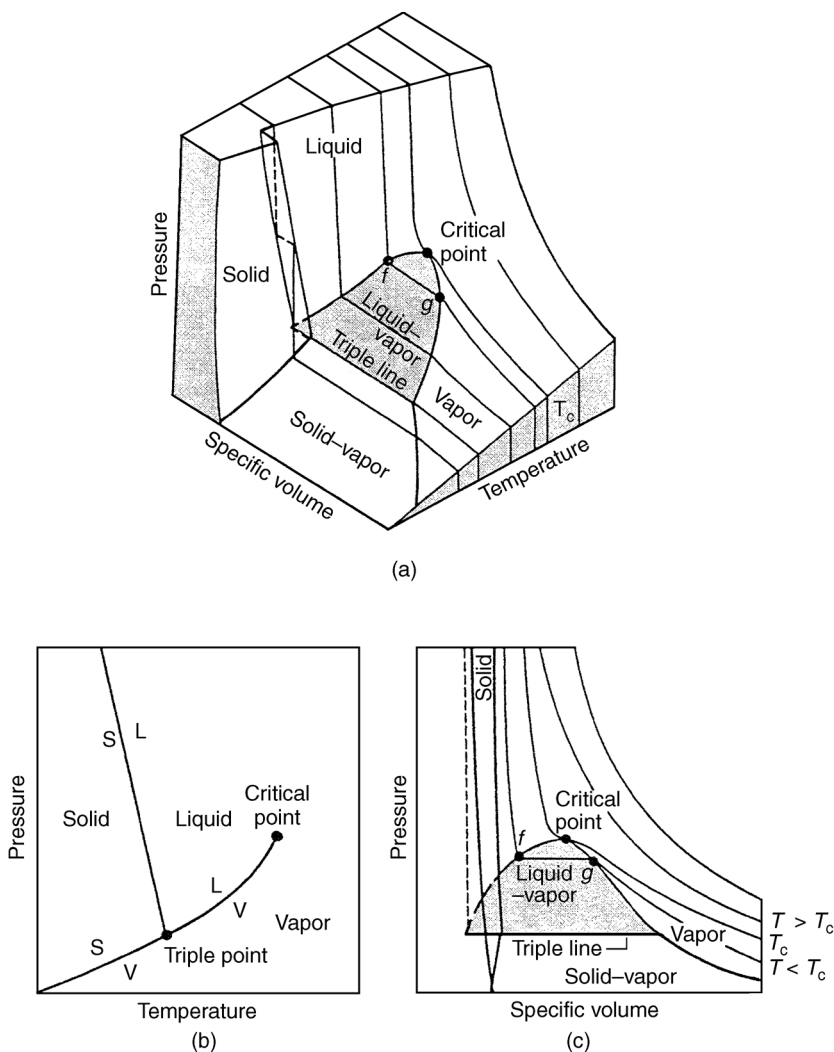



FIGURE 12.1 Pressure-specific volume-temperature surface and projections for water (not to scale).

phase and the two-phase liquid-vapor region is the saturated liquid line. The state denoted by f is a saturated liquid state. The saturated vapor line separates the vapor region and the two-phase liquid-vapor region. The state denoted by g is a saturated vapor state. The saturated liquid line and the saturated vapor line meet at the *critical point*. At the critical point, the pressure is the *critical pressure* p_c, and the temperature is the *critical temperature* T_c. Three phases can coexist in equilibrium along the line labeled *triple line*. The triple line projects onto a point on the phase diagram: the triple point.

When a phase change occurs during constant pressure heating or cooling, the temperature remains constant as long as both phases are present. Accordingly, in the two-phase liquid-vapor region, a line of constant pressure is also a line of constant temperature. For a specified pressure, the corresponding temperature is called the *saturation temperature*. For a specified temperature, the corresponding pressure is called the *saturation pressure*. The region to the right of the saturated vapor line is known as the *superheated vapor region* because the vapor exists at a temperature greater than the saturation temperature for its pressure. The region to the left of the saturated liquid line is known as the *compressed liquid region* because the liquid is at a pressure higher than the saturation pressure for its temperature.

When a mixture of liquid and vapor coexists in equilibrium, the liquid phase is a saturated liquid and the vapor phase is a saturated vapor. The total volume of any such mixture is $V = V_f + V_g$; or, alternatively, $mv = m_f v_f + m_g v_g$, where m and v denote mass and specific volume, respectively. Dividing by the total mass of the mixture m and letting the *mass fraction* of the vapor in the mixture, m_g/m, be symbolized by x, called the *quality*, the apparent specific volume v of the mixture is

$$v = (1-x)v_f + xv_g = v_f + xv_{fg} \tag{12.16a}$$

where $v_{fg} = v_g - v_f$. Expressions similar in form can be written for internal energy, enthalpy, and entropy:

$$u = (1-x)u_f + xu_g = u_f + xu_{fg} \tag{12.16b}$$

$$h = (1-x)h_f + xh_g = h_f + xh_{fg} \tag{12.16c}$$

$$s = (1-x)s_f + xs_g = s_f + xs_{fg} \tag{12.16d}$$

Thermodynamic Data Retrieval

Tabular presentations of pressure, specific volume, and temperature are available for practically important gases and liquids. The tables normally include other properties useful for thermodynamic analyses, such as internal energy, enthalpy, and entropy. The various *steam tables* included in the references of this chapter provide examples. Computer software for retrieving the properties of a wide range of substances is also available, as, for example, the ASME Steam Tables (1993) and Bornakke and Sonntag (1996). Increasingly, textbooks come with computer disks providing thermodynamic property data for water, certain refrigerants, and several gases modeled as ideal gases—see, e.g., Moran and Shapiro (2000).

The sample *steam table data* presented in Table 12.3 are representative of data available for substances commonly encountered in engineering practice. The form of the tables and how they are used are assumed to be familiar. In particular, the use of *linear interpolation* with such tables is assumed known.

Specific internal energy, enthalpy, and entropy data are determined relative to arbitrary datums and such datums vary from substance to substance. Referring to Table 12.3a, the datum state for the specific internal energy and specific entropy of water is seen to correspond to saturated liquid water at 0.01°C (32.02°F), the triple point temperature. The value of each of these properties is set to zero at this state. If calculations are performed involving only differences in a particular specific property, the datum cancels. When there are changes in chemical composition during the process, special care must be exercised. The approach followed when composition changes due to chemical reaction is considered in Moran and Shapiro (2000).

Liquid water data (see Table 12.3d) suggests that at fixed temperature the variation of specific volume, internal energy, and entropy with pressure is slight. The variation of specific enthalpy with pressure at fixed temperature is somewhat greater because pressure is explicit in the definition of enthalpy. This behavior for v, u, s, and h is exhibited generally by liquid data and provides the basis for the following set of equations for estimating property data at liquid states from saturated liquid data:

$$v(T, p) \approx v_f(T) \tag{12.17a}$$

$$u(T, p) \approx u_f(T) \tag{12.17b}$$

$$h(T, p) \approx h_f(T) + v_f[p - p_{sat}(T)] \tag{12.17c}$$

$$s(T, p) \approx s_f(T) \tag{12.17d}$$

The subscript f denotes the saturated liquid state at the temperature T, and p_{sat} is the corresponding saturation pressure. The underlined term of Eq. (12.17c) is usually negligible, giving $h(T,p) \approx h_f(T)$.

Graphical representations of property data also are commonly used. These include the p-T and p-v diagrams of Fig. 12.1, the T-s diagram of Fig. 12.2, the h-s (Mollier) diagram of Fig. 12.3, and the p-h diagram of Fig. 12.4. The compressibility charts considered next use the compressibility factor as one of the coordinates.

Compressibility Charts

The p-v-T relation for a wide range of common gases is illustrated by the generalized compressibility chart of Fig. 12.5. In this chart, the compressibility factor, Z, is plotted vs. the *reduced* pressure, p_R, *reduced* temperature, T_R, and *pseudoreduced* specific volume, v'_R where

$$Z = \frac{p\bar{v}}{\bar{R}T} \tag{12.18}$$

In this expression $\bar{v}$ is the specific volume on a molar basis (m^3/kmol, for example) and $\bar{R}$ is the *universal gas constant* (8314 N · m/kmol · K, for example). The reduced properties are

$$p_R = \frac{p}{p_c}, \qquad T_R = \frac{T}{T_c}, \qquad v'_R = \frac{\bar{v}}{(\bar{R}T_c/p_c)} \tag{12.19}$$

where p_c and T_c denote the critical pressure and temperature, respectively. Values of p_c and T_c are obtainable from the literature—see, for example, Moran and Shapiro (2000). The reduced isotherms of Fig. 12.5 represent the best curves fitted to the data of several gases. For the 30 gases used in developing the chart, the deviation of observed values from those of the chart is at most on the order of 5% and for most ranges is much less.

Analytical Equations of State

Considering the isotherms of Fig. 12.5, it is plausible that the variation of the compressibility factor might be expressed as an equation, at least for certain intervals of p and T. Two expressions can be written that enjoy a theoretical basis. One gives the compressibility factor as an infinite series expansion in pressure,

$$Z = 1 + \hat{B}(T)p + \hat{C}(T)p^2 + \hat{D}(T)p^3 + \cdots \tag{12.20a}$$

and the other is a series in $1/\bar{v}$,

$$Z = 1 + \frac{B(T)}{\bar{v}} + \frac{C(T)}{\bar{v}^2} + \frac{D(T)}{\bar{v}^3} + \cdots \tag{12.20b}$$

Such equations of state are known as *virial expansions*, and the coefficients $\hat{B}$, $\hat{C}$, $\hat{D}$ … and B, C, D… are called *virial coefficients*. In principle, the virial coefficients can be calculated using expressions from statistical mechanics derived from consideration of the force fields around the molecules. Thus far the first few coefficients have been calculated for gases consisting of relatively simple molecules. The coefficients also can be found, in principle, by fitting p-v-T data in particular realms of interest. Only the first few coefficients can be found accurately this way, however, and the result is a *truncated* equation valid only at certain states.

Over 100 equations of state have been developed in an attempt to portray accurately the p-v-T behavior of substances and yet avoid the complexities inherent in a full virial series. In general, these equations exhibit little in the way of fundamental physical significance and are mainly empirical in character. Most are developed for gases, but some describe the p-v-T behavior of the liquid phase, at least qualitatively.

TABLE 12.3 Sample Steam Table Data

(a) Properties of Saturated Water (Liquid-Vapor): Temperature Table

Temp (°C)	Pressure (bar)	Specific Volume (m^3/kg) Saturated Liquid ($v_f \times 10^3$)	Specific Volume (m^3/kg) Saturated Vapor (v_g)	Internal Energy (kJ/kg) Saturated Liquid (u_f)	Internal Energy (kJ/kg) Saturated Vapor (u_g)	Enthalpy (kJ/kg) Saturated Liquid (h_f)	Enthalpy (kJ/kg) Evap. (h_{fg})	Enthalpy (kJ/kg) Saturated Vapor (h_g)	Entropy (kJ/kg · K) Saturated Liquid (s_f)	Entropy (kJ/kg · K) Saturated Vapor (s_g)
.01	0.00611	1.0002	206.136	0.00	2375.3	0.01	2501.3	2501.4	0.0000	9.1562
4	0.00813	1.0001	157.232	16.77	2380.9	16.78	2491.9	2508.7	0.0610	9.0514
5	0.00872	1.0001	147.120	20.97	2382.3	20.98	2489.6	2510.6	0.0761	9.0257
6	0.00935	1.0001	137.734	25.19	2383.6	25.20	2487.2	2512.4	0.0912	9.0003
8	0.01072	1.0002	120.917	33.59	2386.4	33.60	2482.5	2516.1	0.1212	8.9501

(b) Properties of Saturated Water (Liquid-Vapor): Pressure Table

Pressure (bar)	Temp (°C)	Specific Volume (m^3/kg) Saturated Liquid ($v_f \times 10^3$)	Specific Volume (m^3/kg) Saturated Vapor (v_g)	Internal Energy (kJ/kg) Saturated Liquid (u_f)	Internal Energy (kJ/kg) Saturated Vapor (u_g)	Enthalpy (kJ/kg) Saturated Liquid (h_f)	Enthalpy (kJ/kg) Evap. (h_{fg})	Enthalpy (kJ/kg) Saturated Vapor (h_g)	Entropy (kJ/kg · K) Saturated Liquid (s_f)	Entropy (kJ/kg · K) Saturated Vapor (s_g)
0.04	28.96	1.0040	34.800	121.45	2415.2	121.46	2432.9	2554.4	0.4226	8.4746
0.06	36.16	1.0064	23.739	151.53	2425.0	151.53	2415.9	2567.4	0.5210	8.3304
0.08	41.51	1.0084	18.103	173.87	2432.2	173.88	2403.1	2577.0	0.5926	8.2287
0.10	45.81	1.0102	14.674	191.82	2437.9	191.83	2392.8	2584.7	0.6493	8.1502
0.20	60.06	1.0172	7.649	251.38	2456.7	251.40	2358.3	2609.7	0.8320	7.9085

(c) Properties of Superheated Water Vapor

T(°C)	v(m³/kg)	u(kJ/kg)	h(kJ/kg)	s(kJ/kg · K)	v(m³/kg)	u(kJ/kg)	h(kJ/kg)	s(kJ/kg · K)
	p = 0.06 bar = 0.006 MPa (T_{sat} 36.16°C)				p = 0.35 bar = 0.035 MPa (T_{sat} = 72.69°C)			
Sat.	23.739	2425.0	2567.4	8.3304	4.526	2473.0	2631.4	7.7158
80	27.132	2487.3	2650.1	8.5804	4.625	2483.7	2645.6	7.7564
120	30.219	2544.7	2726.0	8.7840	5.163	2542.4	2723.1	7.9644
160	33.302	2602.7	2802.5	8.9693	5.696	2601.2	2800.6	8.1519
200	36.383	2661.4	2879.7	9.1398	6.228	2660.4	2878.4	8.3237

(d) Properties of Compressed Liquid Water

T(°C)	$v \times 10^3$ (m³/kg)	u(kJ/kg)	h(kJ/kg)	s(kJ/kg · K)	$v \times 10^3$ (m³/kg)	u(kJ/kg)	h(kJ/kg)	s(kJ/kg · K)
	p = 25 bar = 2.5 MPa (T_{sat} 223.99°C)				p = 50 bar = 5.0 MPa (T_{sat} = 263.99°C)			
20	1.0006	83.80	86.30	0.2961	0.9995	83.65	88.65	0.2956
80	1.0280	334.29	336.86	1.0737	1.0268	333.72	338.85	1.0720
140	1.0784	587.82	590.52	1.7369	1.0768	586.76	592.15	1.7343
200	1.1555	849.9	852.8	2.3294	1.1530	848.1	853.9	2.3255
Sat.	1.1973	959.1	962.1	2.5546	1.2859	1147.8	1154.2	2.9202

Source: Moran, M.J. and Shapiro, H.N. 2000. *Fundamentals of Engineering Thermodynamics*, 4th ed. Wiley, New York, as extracted from Keenan, J.H., Keyes, F.G., Hill, P.G., and Moore, J.G. 1969. *Steam Tables*. Wiley, New York.

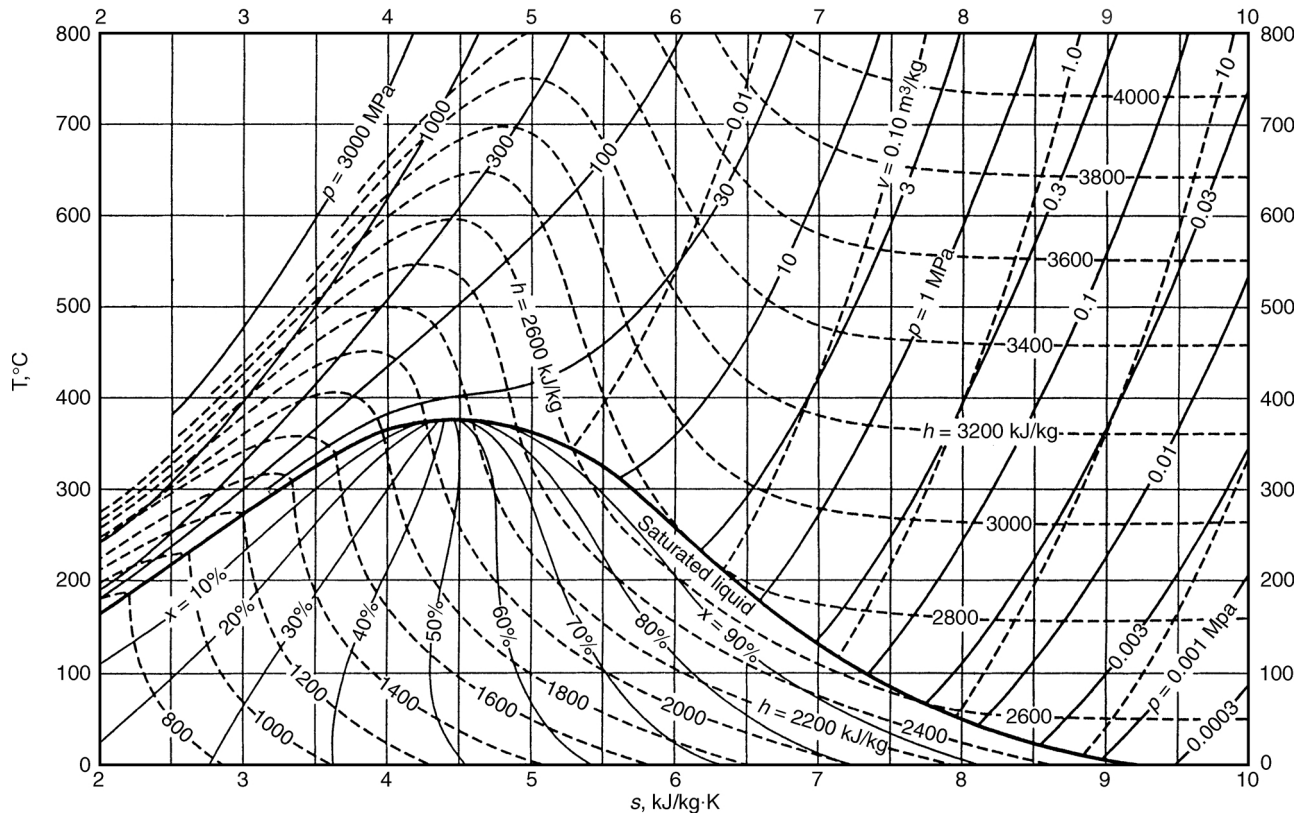



FIGURE 12.2 Temperature-entropy diagram for water. (*Source:* Jones, J.B. and Dugan, R.E. 1996. *Engineering Thermodynamics,* Prentice-Hall, Englewood Cliffs, NJ, based on data and formulations from Haar, L., Gallagher, J.S., and Kell, G.S. 1984. *NBS/NRC Steam Tables.* Hemisphere, Washington, D.C.)

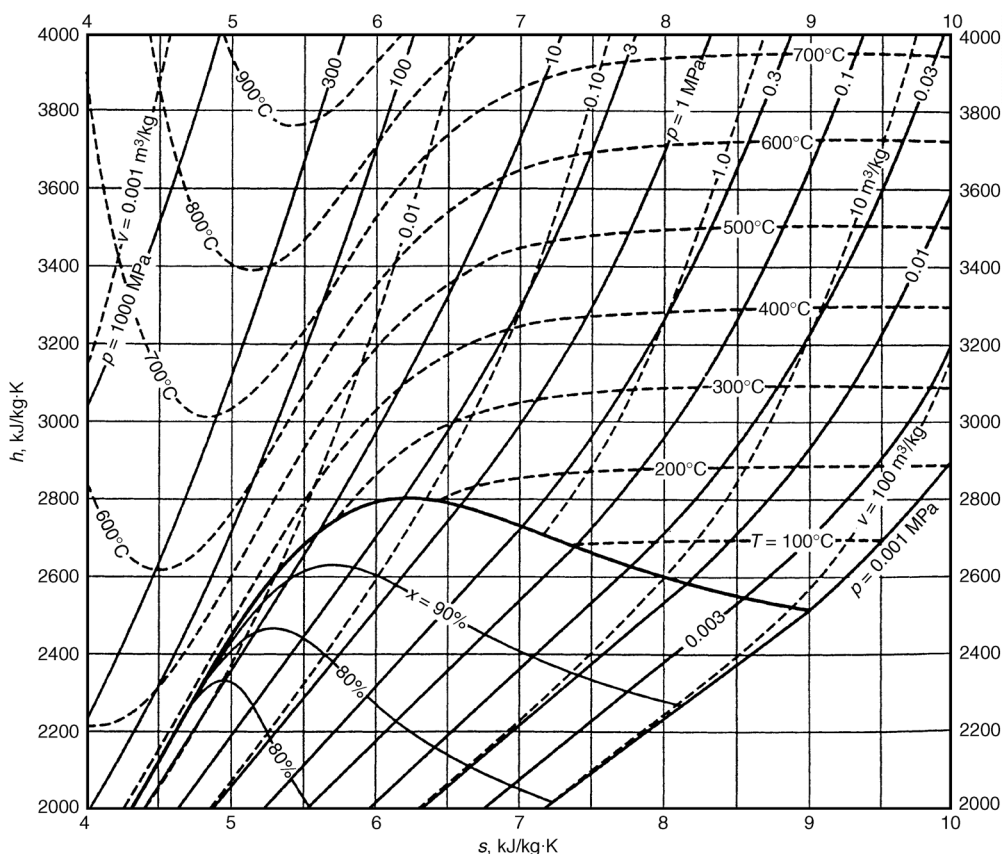



FIGURE 12.3 Enthalpy-entropy (Mollier) diagram for water. (*Source:* Jones, J.B. and Dugan, R.E. 1996. *Engineering Thermodynamics*. Prentice-Hall, Englewood Cliffs, NJ, based on data and formulations from Haar, L., Gallagher, J.S., and Kell, G.S. 1984. *NBS/NRC Steam Tables*. Hemisphere, Washington, D.C.)

Every equation of state is restricted to particular states. The realm of applicability is often indicated by giving an interval of pressure, or density, where the equation can be expected to represent the p-v-T behavior faithfully. For further discussion of equations of state see Reid and Sherwood (1966) and Reid et al. (1987).

Ideal Gas Model

Inspection of the generalized compressibility chart, Fig. 12.5, shows that when p_R is small, and for many states when T_R is large, the value of the compressibility factor Z is close to 1. In other words, for pressures that are low relative to p_c , and for many states with temperatures high relative to T_c, the compressibility factor approaches a value of 1. Within the indicated limits, it may be assumed with reasonable accuracy that $Z = 1$—i.e.,

$$p\bar{v} = \bar{R}T \qquad \text{or} \qquad pv = RT \tag{12.21a}$$

Other forms of this expression in common use are

$$pV = n\bar{R}T, \qquad pV = mRT \tag{12.21b}$$

In these equations, $n = m/\mathcal{M}$, $\bar{v} = \mathcal{M}v$, and the *specific gas constant* is $R = \bar{R}/\mathcal{M}$, where $\mathcal{M}$ denotes the molecular weight.

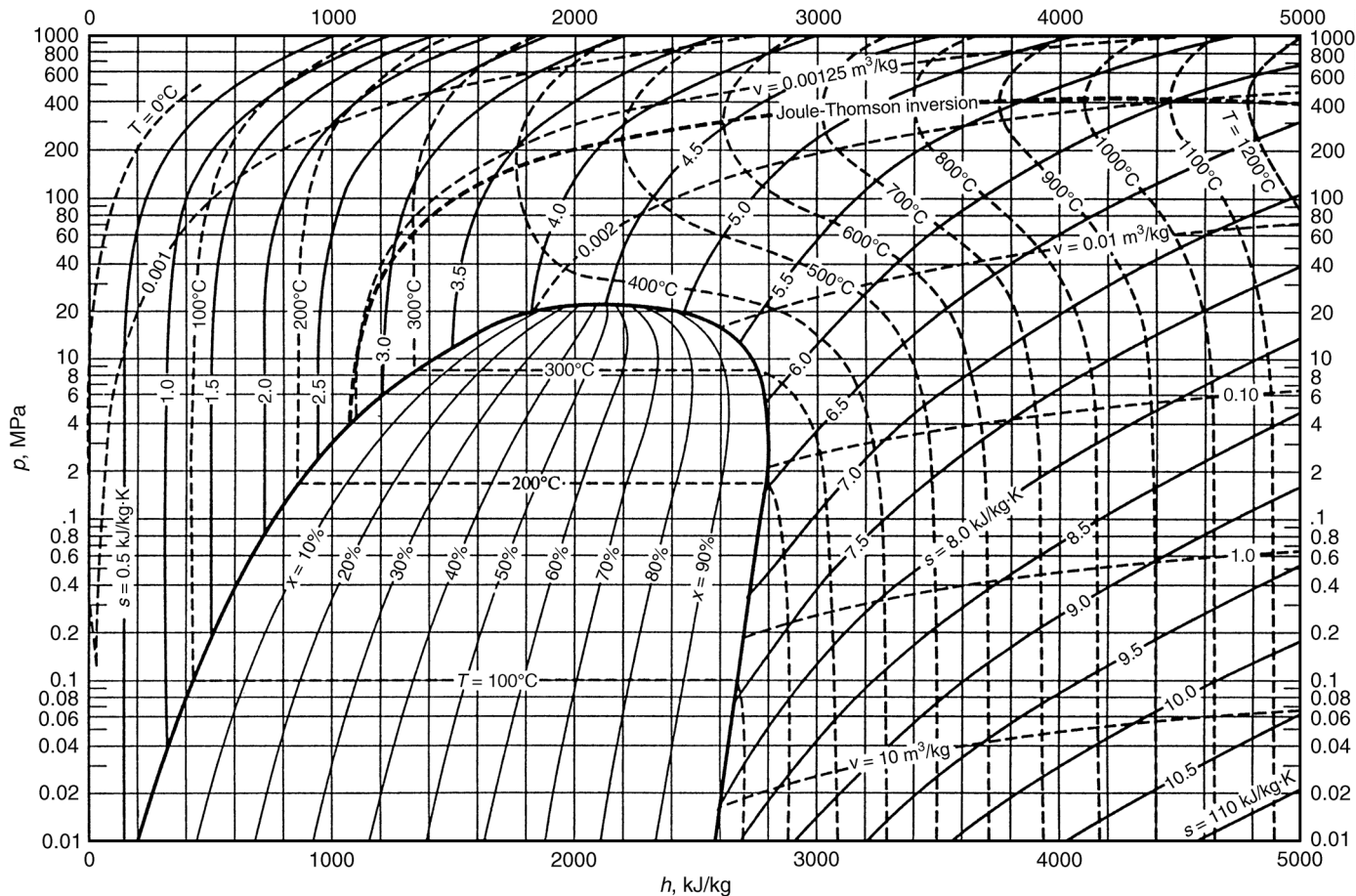



FIGURE 12.4 Pressure-enthalpy diagram for water. (*Source:* Jones, J.B. and Dugan, R.E. 1996. *Engineering Thermodynamics.* Prentice-Hall, Englewood Cliffs, NJ, based on data and formulations from Haar, L., Gallagher, J.S., and Kell, G.S. 1984. *NBS/NRC Steam Tables.* Hemisphere, Washington, D.C.)

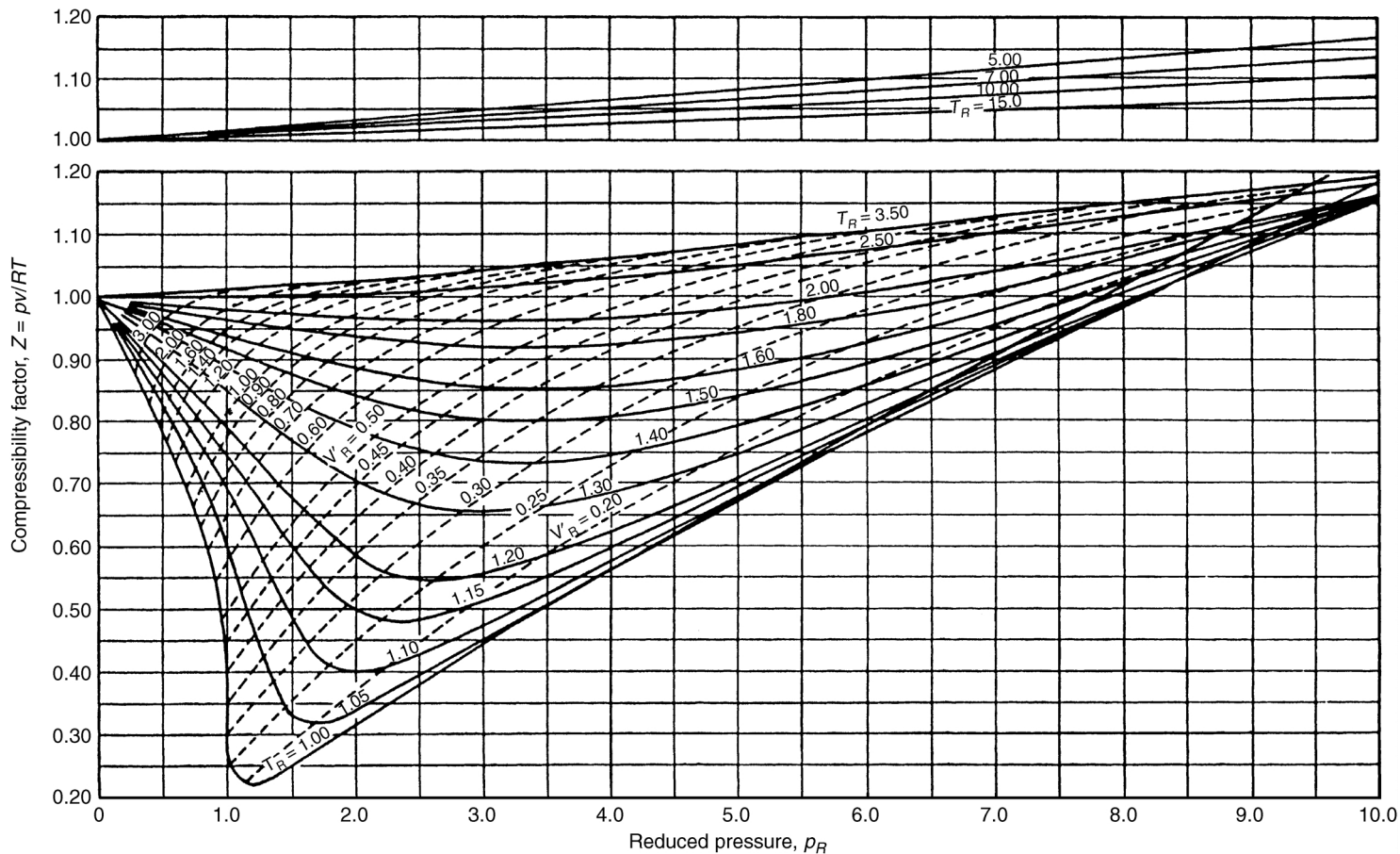



FIGURE 12.5 Generalized compressibility chart ($T_R = T/T_c$, $p_R = p/p_c$, $v'_R = \bar{v}p_c/\bar{R}T_c$) for $p_R \leq 10$. (*Source:* Obert, E.F. 1960 *Concepts of Thermodynamics.* McGraw-Hill, New York.)

TABLE 12.4 Ideal Gas Expressions for Δh, Δu, and Δs

Variable Specific Heats		Constant Specific Heats[b]	
$h(T_2) - h(T_1) = \int_{T_1}^{T_2} c_p(T)\,dT$	(1)	$h(T_2) - h(T_1) = c_p(T_2 - T_1)$	(1′)
$s(T_2, p_2) - s(T_1, p_1) = \int_{T_1}^{T_2} \frac{c_p(T)}{T} dT - R\ln\frac{p_2}{p_1}$	(2)[a]	$s(T_2, p_2) - s(T_1, p_1) = c_p \ln\frac{T_2}{T_1} - R\ln\frac{p_2}{p_1}$	(2′)
$u(T_2) - u(T_1) = \int_{T_1}^{T_2} c_v(T)\,dT$	(3)	$u(T_2) - u(T_1) = c_v(T_2 - T_1)$	(3′)
$s(T_2, v_2) - s(T_1, v_1) = \int_{T_1}^{T_2} \frac{c_v(T)}{T} dT + R\ln\frac{v_2}{v_1}$	(4)	$s(T_2, v_2) - s(T_1, v_1) = c_v \ln\frac{T_2}{T_1} + R\ln\frac{v_2}{v_1}$	(4′)
$s_2 = s_1$		$s_2 = s_1$	
$\frac{p_r(T_2)}{p_r(T_1)} = \frac{p_2}{p_1}$	(5)	$\frac{T_2}{T_1} = \left(\frac{p_2}{p_1}\right)^{(k-1)/k}$	(5′)
$\frac{v_r(T_2)}{v_r(T_1)} = \frac{v_2}{v_1}$	(6)	$\frac{T_2}{T_1} = \left(\frac{v_2}{v_1}\right)^{k-1}$	(6′)

[a] Alternatively, $s(T_2,p_2) - s(T_1,p_1) = s^\circ(T_2) - s^\circ(T_1) - R\ln\frac{p_2}{p_1}$.

[b] c_p and c_v are average values over the temperature interval from T_1 to T_2.

It can be shown that $(\partial u/\partial v)_T$ vanishes identically for a gas whose equation of state is exactly given by Eq. (12.21), and thus the specific internal energy depends only on temperature. This conclusion is supported by experimental observations beginning with the work of Joule, who showed that the internal energy of air at low density depends primarily on temperature.

The above considerations allow for an *ideal gas model* of each real gas: (1) the equation of state is given by Eq. (12.21) and (2) the internal energy, enthalpy, and specific heats (Table 12.2) are functions of temperature alone. The real gas approaches the model in the limit of low reduced pressure. At other states the actual behavior may depart substantially from the predictions of the model. Accordingly, caution should be exercised when invoking the ideal gas model lest error is introduced.

Specific heat data for gases can be obtained by direct measurement. When extrapolated to zero pressure, ideal gas-specific heats result. Ideal gas-specific heats also can be calculated using molecular models of matter together with data from spectroscopic measurements. The following ideal gas-specific heat relations are frequently useful:

$$c_p(T) = c_v(T) + R \tag{12.22a}$$

$$c_p = \frac{kR}{k-1}, \qquad c_v = \frac{R}{k-1} \tag{12.22b}$$

where $k = c_p/c_v$.

For processes of an ideal gas between states 1 and 2, Table 12.4 gives expressions for evaluating the changes in specific enthalpy, Δh, specific entropy, Δs, and specific internal energy, Δu. Relations also are provided for processes of an ideal gas between states having the same specific entropy: $s_2 = s_1$. Property relations and data required by the expressions of Table 12.4: h, u, c_p, c_v, p_r, v_r, and s° are obtainable from the literature—see, for example, Moran and Shapiro (2000).

12.4 Vapor and Gas Power Cycles

Vapor and gas power systems develop electrical or mechanical power from sources of chemical, solar, or nuclear origin. In *vapor* power systems the *working fluid,* normally water, undergoes a phase change from liquid to vapor, and conversely. In *gas* power systems, the working fluid remains a gas throughout, although the composition normally varies owing to the introduction of a fuel and subsequent combustion.

The processes taking place in power systems are sufficiently complicated that idealizations are typically employed to develop tractable thermodynamic models. The *air standard analysis* of gas power systems considered in the present section is a noteworthy example. Depending on the degree of idealization, such models may provide only qualitative information about the performance of the corresponding real-world systems. Yet such information frequently is useful in gauging how changes in major operating parameters might affect actual performance. Elementary thermodynamic models also can provide simple settings to assess, at least approximately, the advantages and disadvantages of features proposed to improve thermodynamic performance.

Work and Heat Transfer in Internally Reversible Processes

Expressions giving work and heat transfer in internally reversible processes are useful in describing the themodynamic performance of vapor and gas cycles. Important special cases are presented in the discussion to follow. For a gas as the system, the work of expansion arises from the force exerted by the system to move the boundary against the resistance offered by the surroundings:

$$W = \int_1^2 F\,dx = \int_1^2 pA\,dx$$

where the force is the product of the moving area and the pressure exerted by the system there. Noting that Adx is the change in total volume of the system,

$$W = \int_1^2 p\,dV$$

This expression for work applies to both actual and internal expansion processes. However, for an internally reversible process p is not only the pressure at the moving boundary but also the pressure throughout the system. Furthermore, for an internally reversible process the volume equals mv, where the specific volume v has a single value throughout the system at a given instant. Accordingly, the work of an internally reversible expansion (or compression) process per unit of system mass is

$$\left(\frac{W}{m}\right)_{\substack{\text{int}\\\text{rev}}} = \int_1^2 p\,dv \tag{12.23}$$

When such a process of a closed system is represented by a continuous curve on a plot of pressure vs. specific volume, the area under the curve is the magnitude of the work per unit of system mass: area a-b-c′-d′ of Fig. 12.6.

For one-inlet, one-exit control volumes in the absence of internal irreversibilities, the following expression gives the work developed per unit of mass flowing:

$$\left(\frac{\dot{W}}{\dot{m}}\right)_{\substack{\text{int}\\\text{rev}}} = -\int_i^e v\,dp + \frac{\mathrm{v}_i^2 - \mathrm{v}_e^2}{2} + g(z_i - z_e) \tag{12.24a}$$

where the integral is performed from inlet to exit (see Moran and Shapiro (2000) for discussion). If there is no significant change in kinetic or potential energy from inlet to exit, Eq. (12.24a) reads

$$\left(\frac{\dot{W}}{\dot{m}}\right)_{\substack{\text{int}\\\text{rev}}} = -\int_i^e v\,dp \quad (\Delta ke = \Delta pe = 0) \tag{12.24b}$$

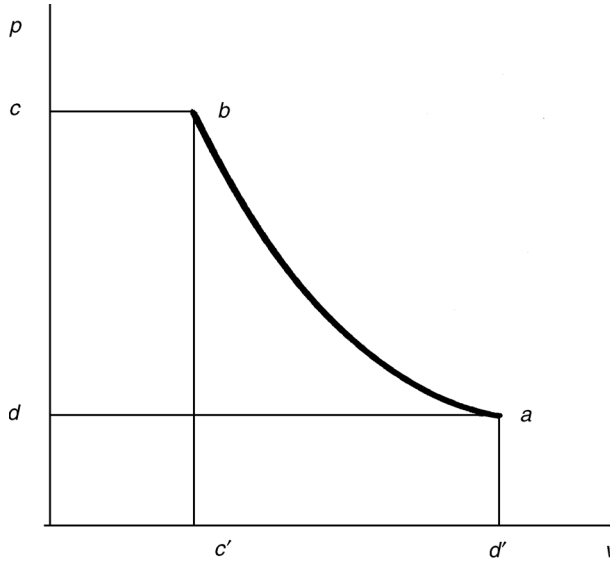



FIGURE 12.6 Internally reversible process on p-v coordinates.

The specific volume remains approximately constant in many applications with liquids. Then Eq. (12.24b) becomes

$$\left(\frac{\dot{W}}{\dot{m}}\right)_{\substack{\text{int}\\\text{rev}}} = -v(p_e - p_i) \quad (v = \text{constant}) \tag{12.24c}$$

When the states visited by a unit of mass flowing without irreversibilities from inlet to outlet are described by a continuous curve on a plot pressure vs. specific volume, as shown in Fig. 12.6, the magnitude of the integral $\int v\,dp$ of Eqs. (12.24a) and (12.24b) is represented by the area a-b-c-d *behind* the curve.

For an internally reversible process of a closed system between state 1 and state 2, the heat transfer per unit of system mass is

$$\left(\frac{Q}{m}\right)_{\substack{\text{int}\\\text{rev}}} = \int_1^2 T\,ds \tag{12.25}$$

For a one-inlet, one-exit control volume in the absence of internal irreversibilities, the following expression gives the heat transfer per unit of mass flowing from inlet i to exit e:

$$\left(\frac{\dot{Q}}{\dot{m}}\right)_{\substack{\text{int}\\\text{rev}}} = \int_i^e T\,ds \tag{12.26}$$

When any such process is represented by a continuous curve on a plot of temperature vs. specific entropy, the area *under* the curve is the magnitude of the heat transfer per unit of mass.

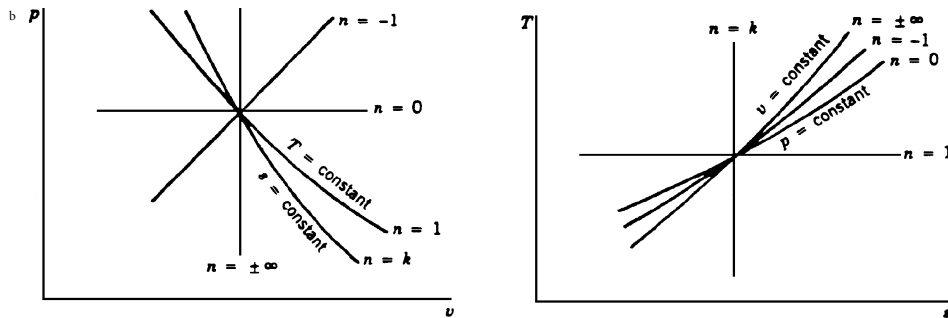
Polytropic Processes

An internally reversible process described by the expression pv^n = constant is called a *polytropic process* and n is the *polytropic exponent*. In certain applications n can be obtained by fitting pressure-specific volume data. Although this expression can be applied when real gases are considered, it most generally appears in practice together with the use of the ideal gas model. Table 12.5 provides several expressions applicable to polytropic processes and the special forms they take when the ideal gas model is assumed. The expressions for $\int p\,dv$ and $\int v\,dp$ have application to work evaluations with Eqs. (12.23) and (12.24), respectively.

TABLE 12.5 Polytropic Processes: pv^n = Constant[a]

General		Ideal Gas[b]	
$\frac{p_2}{p_1} = \left(\frac{v_2}{v_1}\right)^n$	(1)	$\frac{p_2}{p_1} = \left(\frac{v_1}{v_2}\right)^n = \left(\frac{T_2}{T_1}\right)^{n/(n-1)}$	(1′)
$n = 0$: constant pressure $n = \pm\infty$: constant specific volume		$n = 0$: constant pressure $n = \pm\infty$: constant specific volume $n = 1$: constant temperature $n = k$: constant specific entropy when k is constant	
$n = 1$		$n = 1$	
$\int_1^2 p\,dv = p_1 v_1 \ln\frac{v_2}{v_1}$	(2)	$\int_1^2 p\,dv = RT\ln\frac{v_2}{v_1}$	(2′)
$-\int_1^2 v\,dp = -p_1 v_1 \ln\frac{p_2}{p_1}$	(3)	$-\int_1^2 v\,dp = -RT\ln\frac{p_2}{p_1}$	(3′)
$n \neq 1$		$n \neq 1$	
$\int_1^2 p\,dv = \frac{p_2 v_2 - p_1 v_1}{1-n} = \frac{p_1 v_1}{n-1}\left[1-\left(\frac{p_2}{p_1}\right)^{(n-1)/n}\right]$	(4)	$\int_1^2 p\,dv = \frac{R(T_2 - T_1)}{1-n} = \frac{RT_1}{n-1}\left[1-\left(\frac{p_2}{p_1}\right)^{(n-1)/n}\right]$	(4′)
$-\int_1^2 v\,dp = \frac{n}{1-n}(p_2 v_2 - p_1 v_1) = \frac{np_1 v_1}{n-1}\left[1-\left(\frac{p_2}{p_1}\right)^{(n-1)/n}\right]$	(5)	$-\int_1^2 v\,dp = \frac{nR}{1-n}(T_2 - T_1) = \frac{nRT_1}{n-1}\left[1-\left(\frac{p_2}{p_1}\right)^{(n-1)/n}\right]$	(5′)

[a] For polytropic processes of closed systems where volume change is the only work mode, Eqs. (2), (4), and (2′), (4′) are applicable with Eq. (12.23) to evaluate the work. When each unit of mass passing through a one-inlet, one-exit control volume at steady state undergoes a polytropic process, Eqs. (3), (5), and (3′), (5′) are applicable with Eqs. (12.24a) and (12.24b) to evaluate the power. Also note that generally, $-\int_1^2 v\,dp = n\int_1^2 p\,dv$.

[b]

Rankine and Brayton Cycles

In their simplest embodiments vapor power and gas turbine power plants are represented conventionally in terms of four components in series, forming, respectively, the *Rankine cycle* and the *Brayton cycle* shown schematically in Table 12.6. The thermodynamically ideal counterparts of these cycles are composed of four internally reversible processes in series: two isentropic processes alternated with two constant pressure processes. Table 12.6 provides property diagrams of the actual and corresponding ideal cycles. Each actual cycle is denoted 1-2-3-4-1; the ideal cycle is 1-2s-3-4s-1. For simplicity, pressure drops through the boiler, condenser, and heat exchangers are not shown. Invoking Eq. (12.26) for the ideal cycles, the heat added per unit of mass flowing is represented by the area *under* the isobar from state 2s to state 3: area a-2s-3-b-a. The heat rejected is the area *under* the isobar from state 4s to state 1: area

TABLE 12.6 Rankine and Brayton Cycles

Rankine Cycle	Brayton Cycle

$$\left.\begin{matrix}\dot{W}_p \\ \dot{W}_c\end{matrix}\right\} = \dot{m}(h_2 - h_1) \qquad (>0) \qquad (1)$$

$$\dot{Q}_{in} = \dot{m}(h_3 - h_2) \qquad (>0) \qquad (2)$$

$$\dot{W}_t = \dot{m}(h_3 - h_4) \qquad (>0) \qquad (3)$$

$$\dot{Q}_{out} = \dot{m}(h_1 - h_4) \qquad (>0) \qquad (4)$$

a-1-4s-b-a. Enclosed area 1-2s-3-4s-1 represents the net heat added per unit of mass flowing. For any power cycle, the net heat added equals the net work done.

Expressions for the principal energy transfers shown on the schematics of Table 12.6 are provided by Eqs. (1) to (4) of the table. They are obtained by reducing Eq. (12.10a) with the assumptions of negligible heat loss and negligible changes in kinetic and potential energy from the inlet to the exit of each component. All quantities are positive in the directions of the arrows on the figure.

The thermal efficiency of a power cycle is defined as the ratio of the *net* work developed to the total energy added by heat transfer. Using expressions (1)–(3) of Table 12.6, the thermal efficiency is

$$\eta = \frac{(h_3 - h_4) - (h_2 - h_1)}{h_3 - h_2}$$

$$= 1 - \frac{h_4 - h_1}{h_3 - h_2} \qquad (12.27)$$

To obtain the thermal efficiency of the ideal cycle, h_{2s} replaces h_2 and h_{4s} replaces h_4 in Eq. (12.27).

Decisions concerning cycle operating conditions normally recognize that the thermal efficiency tends to increase as the average temperature of heat addition increases and/or the temperature of heat rejection decreases. In the Rankine cycle, a high average temperature of heat addition can be achieved by superheating the vapor prior to entering the turbine and/or by operating at an elevated steam-generator pressure. In the Brayton cycle an increase in the compressor pressure ratio p_2/p_1 tends to increase the average temperature of heat addition. Owing to materials limitations at elevated temperatures and pressures, the state of the working fluid at the turbine inlet must observe practical limits, however. The turbine inlet temperature of the Brayton cycle, for example, is controlled by providing air far in excess of what is required for combustion. In a Rankine cycle using water as the working fluid, a low temperature of heat rejection is typically achieved by operating the condenser at a pressure below 1 atm. To reduce erosion and wear by liquid droplets on the blades of the Rankine cycle steam turbine, at least 90% steam quality should be maintained at the turbine exit: $x_4 > 0.9$.

The back work ratio, bwr, is the ratio of the work required by the pump or compressor to the work developed by the turbine:

$$\text{bwr} = \frac{h_2 - h_1}{h_3 - h_4} \tag{12.28}$$

As a relatively high specific volume vapor expands through the turbine of the Rankine cycle and a much lower specific volume liquid is pumped, the back work ratio is characteristically quite low in vapor power plants—in many cases on the order of 1–2%. In the Brayton cycle, however, both the turbine and compressor handle a relatively high specific volume gas, and the back ratio is much larger, typically 40% or more.

The effect of friction and other irreversibilities for flow through turbines, compressors, and pumps is commonly accounted for by an appropriate *isentropic efficiency*. Referring to Table 12.6 for the states, the isentropic turbine efficiency is

$$\eta_t = \frac{h_3 - h_4}{h_3 - h_{4s}} \tag{12.29a}$$

The isentropic compressor efficiency is

$$\eta_c = \frac{h_{2s} - h_1}{h_2 - h_1} \tag{12.29b}$$

In the isentropic pump efficiency, η_p, which takes the same form as Eq. (12.29b), the numerator is frequently approximated via Eq. (12.24c) as $h_{2s} - h_1 \approx v_1 \Delta p$, where Δp is the pressure rise across the pump.

Simple gas turbine power plants differ from the Brayton cycle model in significant respects. In actual operation, excess air is continuously drawn into the compressor, where it is compressed to a higher pressure; then fuel is introduced and combustion occurs; finally the mixture of combustion products and air expands through the turbine and is subsequently discharged to the surroundings. Accordingly, the low-temperature heat exchanger shown by a dashed line in the Brayton cycle schematic of Table 12.6 is not an actual component, but included only to account formally for the cooling in the surroundings of the hot gas discharged from the turbine.

Another frequently employed idealization used with gas turbine power plants is that of an *air-standard analysis*. An air-standard analysis involves two major assumptions: (1) As shown by the Brayton cycle schematic of Table 12.6, the temperature rise that would be brought about by combustion is effected instead by a heat transfer from an external source. (2) The working fluid throughout the cycle is air, which behaves as an ideal gas. In a cold air-standard analysis the specific heat ratio k for air is taken as constant. Equations (1) to (6) of Table 12.4 apply generally to air-standard analyses. Equations (1′) to (6′)

of Table 12.4 apply to *cold* air-standard analyses, as does the following expression for the turbine power obtained from Table 12.1 (Eq. (10c″)):

$$\dot{W}_t = \dot{m}\frac{kRT_3}{k-1}[1-(p_4/p_3)^{(k-1)/k}] \tag{12.30}$$

An expression similar in form can be written for the power required by the compressor.

Otto, Diesel, and Dual Cycles

Although most gas turbines are also internal combustion engines, the name is usually reserved to *reciprocating* internal combustion engines of the type commonly used in automobiles, trucks, and buses. Two principal types of reciprocating internal combustion engines are the spark-ignition engine and the compression-ignition engine. In a *spark-ignition* engine a mixture of fuel and air is ignited by a spark plug. In a *compression ignition* engine air is compressed to a high-enough pressure and temperature that combustion occurs spontaneously when fuel is injected.

In a *four-stroke* internal combustion engine, a piston executes four distinct strokes within a cylinder for every two revolutions of the crankshaft. Figure 12.7 gives a pressure-displacement diagram as it might be displayed electronically. With the intake valve open, the piston makes an *intake stroke* to draw a fresh charge into the cylinder. Next, with both valves closed, the piston undergoes a *compression stroke* raising the temperature and pressure of the charge. A combustion process is then initiated, resulting in a high-pressure, high-temperature gas mixture. A *power* stroke follows the compression stroke, during which the gas mixture expands and work is done on the piston. The piston then executes an *exhaust stroke* in which the burned gases are purged from the cylinder through the open exhaust valve. Smaller engines operate on *two-stroke* cycles. In two-stroke engines, the intake, compression, expansion, and exhaust operations are accomplished in one revolution of the crankshaft. Although internal combustion engines undergo *mechanical* cycles, the cylinder contents do not execute a *thermodynamic* cycle, since matter is introduced with one composition and is later discharged at a different composition.

A parameter used to describe the performance of reciprocating piston engines is the *mean effective pressure*, or mep. The mean effective pressure is the theoretical constant pressure that, if it acted on the

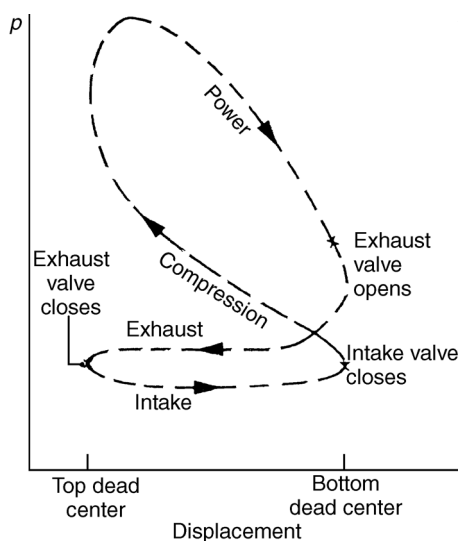



FIGURE 12.7 Pressure-displacement diagram for a reciprocating internal combustion engine.

piston during the power stroke, would produce the same net work as actually developed in one cycle. That is,

$$\text{mep} = \frac{\text{net work for one cycle}}{\text{displacement volume}} \tag{12.31}$$

where the displacement volume is the volume swept out by the piston as it moves from the top dead center to the bottom dead center. For two engines of equal displacement volume, the one with a higher mean effective pressure would produce the greater net work and, if the engines run at the same speed, greater power.

Detailed studies of the performance of reciprocating internal combustion engines may take into account many features, including the combustion process occurring within the cylinder and the effects of irreversibilities associated with friction and with pressure and temperature gradients. Heat transfer between the gases in the cylinder and the cylinder walls and the work required to charge the cylinder and exhaust the products of combustion also might be considered. Owing to these complexities, accurate modeling of reciprocating internal combustion engines normally involves computer simulation.

To conduct *elementary* thermodynamic analyses of internal combustion engines, considerable simplification is required. A procedure that allows engines to be studied *qualitatively* is to employ an *air-standard analysis* having the following elements: (1) a fixed amount of air modeled as an ideal gas is the system; (2) the combustion process is replaced by a heat transfer from an external source and represented in terms of elementary thermodynamic processes; (3) there are no exhaust and intake processes as in an actual engine: the cycle is completed by a constant-volume heat rejection process; (4) all processes are internally reversible.

The processes employed in air-standard analyses of internal combustion engines are selected to represent the events taking place within the engine simply and mimic the appearance of observed pressure-displacement diagrams. In addition to the constant volume heat rejection noted previously, the compression stroke and at least a portion of the power stroke are conventionally taken as isentropic. The heat addition is normally considered to occur at constant volume, at constant pressure, or at constant volume followed by a constant pressure process, yielding, respectively, the Otto, Diesel, and Dual cycles shown in Table 12.7.

Reducing the closed system energy balance, Eq. (12.7b), gives the following expressions for work and heat applicable in each case shown in Table 12.7:

$$\frac{W_{12}}{m} = u_1 - u_2, \qquad \frac{W_{34}}{m} = u_3 - u_4, \qquad \frac{Q_{41}}{m} = u_1 - u_4 \tag{12.32}$$

Table 12.7 provides additional expressions for work, heat transfer, and thermal efficiency identified with each case individually. All expressions for work and heat adhere to the respective sign conventions of Eq. (12.7b). Equations (1) to (6) of Table 12.4 apply generally to air-standard analyses. In a cold air-standard analysis the specific heat ratio k for air is taken as constant. Equations (1′) to (6′) of Table 12.4 apply to cold air-standard analyses, as does Eq. (4′) of Table 12.5, with $n = k$ for the isentropic processes of these cycles.

Referring to Table 12.7, the ratio of specific volumes v_1/v_2 is the *compression ratio, r*. For the Diesel cycle, the ratio v_3/v_2 is the cutoff ratio, r_c. Figure 12.8 shows the variation of the thermal efficiency with compression ratio for an Otto cycle and Diesel cycles having cutoff ratios of 2 and 3. The curves are determined on a cold air-standard basis with $k = 1.4$ using the following expression:

$$\eta = 1 - \frac{1}{r^{k-1}}\left[\frac{r_c^k - 1}{k(r_c - 1)}\right] \quad \text{(constant } k\text{)} \tag{12.33}$$

where the Otto cycle corresponds to $r_c = 1$.

TABLE 12.7 Otto, Diesel, and Dual Cycles

(a) Otto Cycle

T
3
2
4
1
v = c
v = c
s

$$\frac{W_{23}}{m} = 0$$

$$\frac{Q_{23}}{m} = u_3 - u_2$$

$$\eta = 1 - \frac{u_4 - u_1}{u_3 - u_2}$$

(b) Diesel Cycle

$$\frac{W_{23}}{m} = p_2(v_3 - v_2)$$

$$\frac{Q_{23}}{m} = h_3 - h_2$$

$$\eta = 1 - \frac{u_4 - u_1}{h_3 - h_2}$$

(c) Dual cycle

$$\frac{W_{2x}}{m} = 0, \quad \frac{Q_{2x}}{m} = u_x - u_2$$

$$\frac{W_{x3}}{m} = p_3(v_3 - v_2), \quad \frac{Q_{x3}}{m} = h_3 - h_x$$

$$\eta = 1 - \frac{u_4 - u_1}{(u_x - u_2) + (h_3 - h_x)}$$

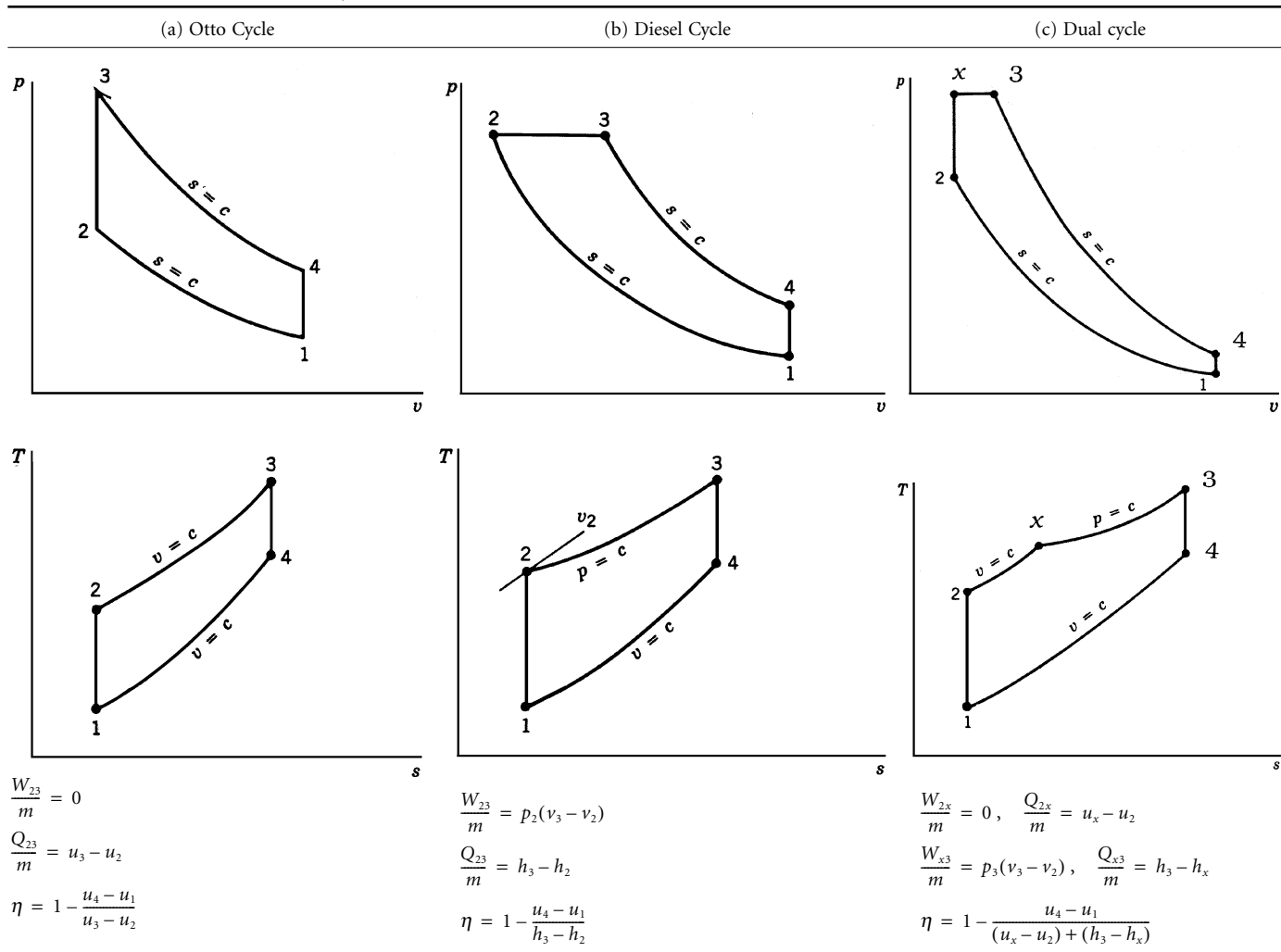
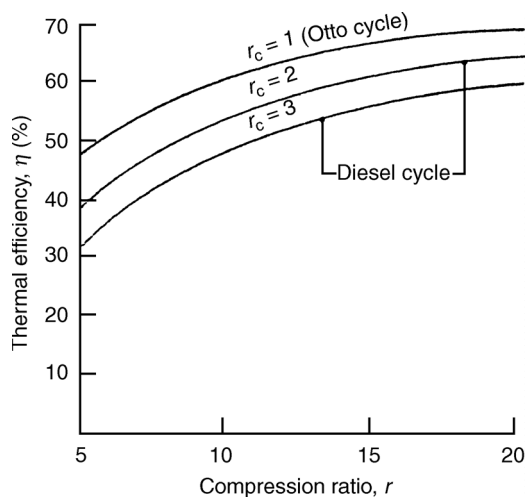



FIGURE 12.8 Thermal efficiency of the cold air-standard Otto and Diesel cycles, $k = 1.4$.

As all processes are internally reversible, areas on the p-v and T-s diagrams of Table 12.7 can be interpreted, respectively, as work and heat transfer. Invoking Eq. (12.23) and referring to the p-v diagrams, the areas under process 3-4 of the Otto cycle, process 2-3-4 of the Diesel cycle, and process x-3-4 of the Dual cycle represent the work done by the gas during the power stroke, per unit of mass. For each cycle, the area under the isentropic process 1-2 represents the work done on the gas during the compression stroke, per unit of mass. The enclosed area of each cycle represents the net work done per unit mass. With Eq. (12.25) and referring to the T-s diagrams, the areas under process 2-3 of the Otto and Diesel cycles and under process 2-x-3 of the Dual cycle represent the heat added per unit of mass. For each cycle, the area under the process 4-1 represents the heat rejected per unit of mass. The enclosed area of each cycle represents the net heat added, which equals the net work done, each per unit of mass.

References

ASHRAE Handbook 1993 Fundamentals. 1993. American Society of Heating, Refrigerating, and Air Conditioning Engineers, Atlanta.

ASME Steam Tables, 6th ed., 1993. ASME Press, Fairfield, NJ.

Bejan, A., Tsatsaronis, G., and Moran, M. 1996. *Thermal Design and Optimization,* John Wiley & Sons, New York.

Bird, R.B., Stewart, W.E., and Lightfoot, E.N. 1960. *Transport Phenomena,* John Wiley & Sons, New York.

Bolz, R.E. and Tuve, G.L. (Eds.). 1973. *Handbook of Tables for Applied Engineering Science,* 2nd ed., CRC Press, Boca Raton, FL.

Bornakke, C. and Sonntag, R.E. 1996. *Tables of Thermodynamic and Transport Properties,* John Wiley & Sons, New York.

Gray, D.E. (Ed.). 1972. *American Institute of Physics Handbook,* McGraw-Hill, New York.

Haar, L., Gallagher, J.S., and Kell, G.S. 1984. *NBS/NRC Steam Tables,* Hemisphere, New York.

Handbook of Chemistry and Physics, annual editions, CRC Press, Boca Raton, FL.

JANAF Thermochemical Tables, 3rd ed., 1986. American Chemical Society and the American Institute of Physics for the National Bureau of Standards.

Jones, J.B. and Dugan, R.E. 1996. *Engineering Thermodynamics,* Prentice-Hall, Englewood Cliffs, NJ.

Keenan, J.H., Keyes, F.G., Hill, P.G., and Moore, J.G. 1969 and 1978. *Steam Tables,* John Wiley & Sons, New York (1969, English Units; 1978, SI Units).

Keenan, J.H., Chao, J., and Kaye, J. 1980 and 1983. *Gas Tables—International Version,* 2nd ed., John Wiley & Sons, New York (1980, English Units; 1983, SI Units).

Knacke, O., Kubaschewski, O., and Hesselmann, K. 1991. *Thermochemical Properties of Inorganic Substances,* 2nd ed., Springer-Verlag, Berlin.

Kotas, T.J. 1995. *The Exergy Method of Thermal Plant Analysis,* Krieger, Melbourne, FL.

Liley, P.E. 1987. *Thermodynamic Properties of Substances,* In *Marks' Standard Handbook for Mechanical Engineers,* E.A. Avallone and T. Baumeister (Eds.), 9th ed., McGraw-Hill, New York, Sec. 4.2.

Liley, P.E., Reid, R.C., and Buck, E. 1984. Physical and chemical data. In *Perrys' Chemical Engineers, Handbook,* R.H. Perry and D.W. Green (Eds.), 6th ed., McGraw-Hill, New York, Sec. 3.

Moran, M.J. 1989. *Availability Analysis—A Guide to Efficient Energy Use,* ASME Press, New York.

Moran, M.J. 1998. Engineering Thermodynamics. In *The CRC Handbook of Mechanical Engineering,* F. Kreith (Ed.), CRC Press, Boca Raton, FL, Chap. 2.

Moran, M.J. and Shapiro, H.N. 2000. *Fundamentals of Engineering Thermodynamics,* 4th ed., John Wiley & Sons, New York.

Moran, M.J. and Shapiro, H.N. 2000. *IT: Interactive Thermodynamics,* Computer Software to Accompany Fundamentals of Engineering Thermodynamics, 4th ed., Intellipro, John Wiley & Sons, New York.

Moran, M.J. and Tsatsaronis, G. 2000. Engineering Thermodynamics. In *The CRC Handbook of Thermal Engineering,* F. Kreith (Ed.), CRC Press, Boca Raton, FL, Chap. 1.

Obert, E.F. 1960. *Concepts of Thermodynamics,* McGraw-Hill, New York.

Preston-Thomas, H. 1990. The International Temperature Scale of 1990 (ITS-90). *Metrologia.* 27: 3–10.

Reid, R.C. and Sherwood, T.K. 1966. *The Properties of Gases and Liquids,* 2nd ed., McGraw-Hill, New York.

Reid, R.C., Prausnitz, J.M., and Poling, B.E. 1987. *The Properties of Gases and Liquids,* 4th ed., McGraw-Hill, New York.

Reynolds, W.C. 1979. *Thermodynamic Properties in SI—Graphs, Tables and Computational Equations for 40 Substances.* Department of Mechanical Engineering, Stanford University, Palo Alto, CA.

Stephan, K. 1994. Tables. In *Dubbel Handbook of Mechanical Engineering,* W. Beitz and K. H. Kuttner (Eds.), Springer-Verlag, London, Sec. C11.

Szargut, J., Morris, D.R., and Steward, F.R. 1988. *Exergy Analysis of Thermal, Chemical and Metallurgical Processes,* Hemisphere, New York.

Van Wylen, G.J., Sonntag, R.E., and Bornakke, C. 1994. *Fundamentals of Classical Thermodynamics,* 4th ed., John Wiley & Sons, New York.

Zemansky, M.W. 1972. *Thermodynamic Symbols, Definitions, and Equations*. In American Institute of Physics Handbook, D.E. Gray (Ed.), McGraw-Hill, New York, Sec. 4b.

13 Modeling and Simulation for MEMS

Carla Purdy
University of Cincinnati

13.1 Introduction

Accurate modeling and efficient simulation, in support of greatly reduced development cycle time and cost, are well established techniques in the miniaturized world of integrated circuits (ICs). Simulation accuracies of 5% or less for parameters of interest are achieved fairly regularly [1], although even much less accurate simulations (25–30%, e.g.) can still be used to obtain valuable information [2]. In the IC world, simulation can be used to predict the performance of a design, to analyze an already existing component, or to support automated synthesis of a design. Eventually, MEMS simulation environments should also be capable of these three modes of operation. The MEMS developer is, of course, most interested in quick access to particular techniques and tools to support the system currently under development. In the long run, however, consistently achieving acceptably accurate MEMS simulations will depend both on the ability of the CAD (computer-aided design) community to develop robust, efficient, user-friendly tools which will be widely available both to cutting-edge researchers and to production engineers and on the existence of readily accessible standardized processes. In this chapter we focus on fundamental approaches which will eventually lead to successful MEMS simulations becoming routine.

We also survey available tools which a MEMS developer can use to achieve good simulation results. Many of these tools build MEMS development systems on platforms already in existence for other technologies, thus leveraging the extensive resources which have gone into previous development and avoiding "reinventing the wheel."

For our discussion of modeling and simulation, the salient characteristics of MEMS are:

1. inclusion and interaction of multiple domains and technologies,
2. both two- and three-dimensional behaviors,
3. mixed digital (discrete) and analog (continuous) input, output, and signals, and
4. micro- (or nano-) scale feature sizes.

Techniques for the manufacture of reliable (two-dimensional) systems with micro- or nano-scale feature sizes (Characteristic 4) are very mature in the field of microelectronics, and it is logical to attempt to extend these techniques to MEMS, while incorporating necessary changes to deal with Characteristics 1–3. Here we survey some of the major principles which have made microelectronics such a rapidly evolving field, and we look at microelectronics tools which can be used or adapted to allow us to apply these principles to MEMS. We also discuss why applying such strategies to MEMS may not always be possible.

13.2 The Digital Circuit Development Process: Modeling and Simulating Systems with Micro- (or Nano-) Scale Feature Sizes

A typical VLSI digital circuit or system process flow is shown in Fig. 13.1, where the dotted lines show the most optimistic point to which the developer must return if errors are discovered. Option A, for a "mature" technology, is supported by efficient and accurate simulators, so that even the first actual implementation ("first silicon") may have acceptable performance. As a process matures, the goal is to have better and better simulations, with a correspondingly smaller chance of discovering major performance flaws after implementation. However, development of models and simulators to support this goal is in itself a major task. Option B (immature technology), at its extreme, would represent an experimental technology for which not enough data are available to support even moderately robust simulations. In modern software and hardware development systems, the emphasis is on tools which provide increasingly good support for the initial stages of this process. This increases the probability that conceptual or design errors will be identified and modifications made as early in the process as possible and thus decreases both development time and overall development cost.

At the microlevel, the development cycle represented by Option A is routinely achieved today for many digital circuits. In fact, the entire process can in some cases be highly automated, so that we have "silicon compilers" or "computers designing computers." Thus, not only design analysis, but even design synthesis is possible. This would be the case for well-established silicon-based CMOS technologies, for example. There are many characteristics of digital systems which make this possible. These include:

- Existence of a small set of basic digital circuit elements. All Boolean functions can be realized by combinations of the logic functions AND, OR, NOT. In fact, all Boolean functions can be realized by combinations of just one gate, a NAND (NOT-AND) gate. So if a "model library" of basic gates (and a few other useful parts, such as I/O pins, multiplexors, and flip-flops) is developed, systems can be implemented just by combining suitable library elements.
- A small set of standardized and well-understood technologies, with well-characterized fabrication processes that are widely available. For example, in the United States, the MOSIS service [3] provides access to a range of such technologies. Similar services elsewhere include CMP in France [4], Europractice in Europe [5], VDEC in Japan [6], and CMC in Canada [7].
- A well-developed educational infrastructure and prototyping facilities. These are provided by all of the services listed above. These types of organization and educational support had their origins in the work of Mead and Conway [8] and continue to produce increasingly sophisticated VLSI engineers.

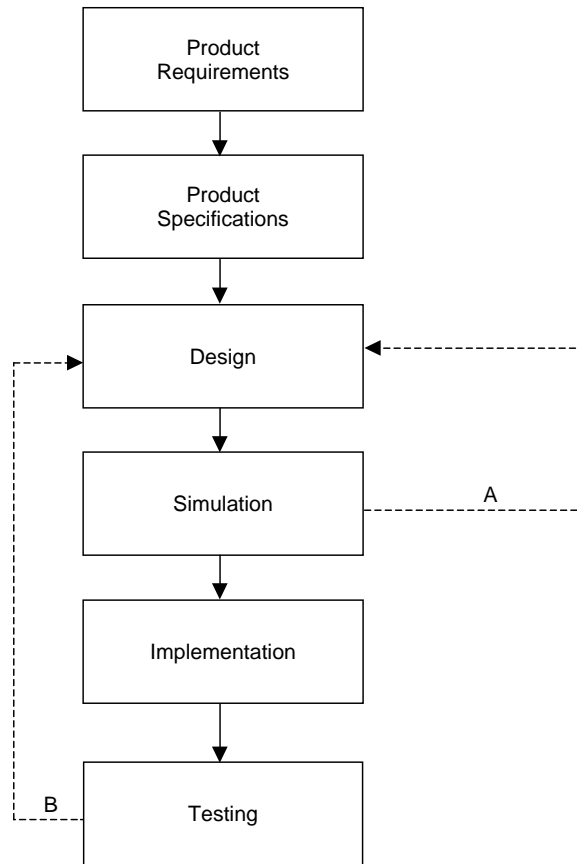



FIGURE 13.1 Product design process. A: mature technology, B: immature technology.

An important aspect of this infrastructure is that it also provides, at relatively low cost, access to example devices and systems, made with stable fabrication processes, whose behavior can be tested and compared to simulation results, thereby enabling improvements in simulation techniques.

- "Levels and views" (abstraction and encapsulation or "information hiding") (see [9]). This concept is illustrated in Fig. 13.2(a). For the VLSI domain, we can identify at least five useful levels of abstraction, from the lowest (layout geometry) to the highest (system specification). We can also "view" a system behaviorally, structurally, or physically. In the *behavioral domain* we describe the functionality of the circuit without specifying how this functionality will be achieved. This allows us to think clearly about what the system needs to do, what inputs are needed, and what outputs will be provided. Thus we can view the component as a "black box" that has specified responses to given inputs. The current through a MOS field effect transistor (MOSFET), given as a function of the gate voltage, is a (low-level) behavioral description, for example. In the *physical domain* we specify the actual physical parts of the circuit. At the lowest levels in this domain, we must choose what material each piece of the circuit will be made from (for example, which pieces of wire will lie in each of the metal layers usually provided in a CMOS circuit) and exactly where each piece will be placed in the actual physical layout. The physical description will be translated directly into mask layouts for the circuit. The *structural domain* is intermediate between physical and behavioral. It provides an interface between the functionality specified in the behavioral domain, which ignores geometry, and the geometry specified in the physical domain, which ignores functionality. In this intermediate domain, we can carry out logic optimization and state minimization, for example.

(a)

Levels	Views		
	Behavioral	Structural	Physical
4	Performance Specifications	CPUs, Memory, Switches, Controllers, Buses	Physical Partitions
3	Algorithms	Modules, Data Structures	Clusters
2	Register Transfers	ALUs, MUXs, Registers	Floorplans
1	Boolean Equations, FSMs	Gates, Flip-flops	Cells, Modules
0	Transfer Functions, Timing	Transistors, Wires, Contacts, Vias	Layout Geometry

(b)

Levels	Views		
	Behavioral	Structural	Physical
4	Performance Specifications	Sensors, Actuators, Systems	Physical Partitions
3		Multiple Energy Domain Components	Clusters
2		Domain-Domain Components	Floorplans
1		Single Energy Domain Components	Cells, Modules
0	Transfer Functions, Timing	Beams, Membranes, Holes, Grooves, Joints	Layout Geometry

FIGURE 13.2 A taxonomy for component development ("levels and views"): (a) standard VLSI classifications, (b) a partial classification for MEMS components.

A schematic diagram is an example of a structural description. Of course, not all circuit characteristics can be completely encapsulated in a single one of these views. For example, if we change the physical size of a wire, we will probably affect the timing, which is a behavioral property. The principle of encapsulation leads naturally to the development of extensive IP (intellectual property), i.e., libraries of increasingly sophisticated components that can be used as "black boxes" by the system developer.

- Well-developed models for basic elements that clearly delineate effects due to changes in design, fabrication process, or environment. For example, in [10], the factors in the basic first-order equations for I_{ds}, the drain-to-source current in an NMOS transistor, can clearly be divided into those under the control of the designer (W/L, the width-to-length ratio for the transistor channel), those dependent on the fabrication process (ε, the permittivity of the gate insulator, and t_{ox}, the thickness of the gate insulator), those dependent on environmental factors (V_{ds} and V_{gs}, the drain-to-source and gate-to-source voltages, respectively), and those that are a function of both the fabrication process and the environment (μ, the effective surface mobility of the carriers in the channel, and V_t, the threshold voltage). More detailed information on modeling MOSFETs can be found in [11]. Identification of fundamental parameters in one stage of the development process can be of great value in other stages. For example, the minimum feature size λ for a given technology can be used to develop a set of "design rules" that express mandatory overlaps and spacings for the different physical materials. A design tool can then be developed to "enforce" these rules, and the consequences can be used to simplify, to some extent, the modeling and simulation stages. The parameter λ can also be used to express effects due to scaling when scaling is valid.

- Mature tools for design and simulation, which have evolved over many generations and for which moderately priced versions are available from multiple sources. For example, many of today's tools incorporate versions of the design tool MAGIC [12] and the simulator SPICE (Simulation Program with Integrated Circuit Emphasis) [13], both of which were originally developed at the University of California, Berkeley. Versions of the SPICE simulator typically support several device models (currently, for example, six or more different MOS models and five different transmission line models), so that a developer can choose the level of device detail appropriate to the task at hand. Free or low-cost versions of both MAGIC and SPICE, as well as extended versions of both tools, are widely available. Many different techniques, such as model binning (optimizing models for specific ranges of model parameters) and inclusion of proprietary process information, are employed to produce better models and simulation results, especially in the HSPICE version of SPICE and in other high-end versions of these tools [11].
- Integrated development systems that are widely available and that provide support for a variety of levels and views, extensive component libraries, user-friendly interfaces and online help, as well as automatic translation between domains, along with error and constraint checking. In an integrated VLSI development system, sophisticated models, simulators, and translators keep track of circuit information for multiple levels and views, while allowing the developer to focus on one level or view at a time. Many development systems available today also support, at the higher levels of abstraction, structured "programming" languages such as VHDL (Very Large Scale Integrated Circuit Hardware Description Language) [14,15] or Verilog [16].

A digital circuit developer has many options, depending on performance constraints, number of units to be produced, desired cost, available development time, etc. At one extreme the designer may choose to develop a "custom" circuit, creating layout geometries, sizing individual transistors, modeling RC effects in individual wires, and validating design choices through extensive low-level SPICE-based simulations. At the other extreme, the developer can choose to produce a PLD (programmable logic device), with a predetermined basic layout geometry consisting of cells incorporating programmable logic and storage (Fig. 13.3) that can be connected as needed to produce the desired device functionality. A high end PLD may contain as many as 100,000 (100 K) cells similar to the one in Fig. 13.3 and an additional 100 K bytes of RAM (random access memory) storage. In an integrated development system, such as those

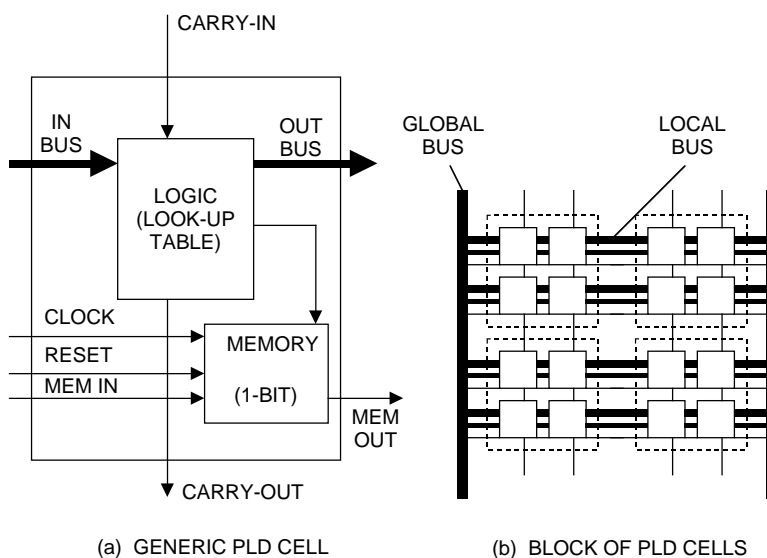



FIGURE 13.3 A generic programmable logic device architecture.

provided by [17] and [18], the developer enters the design in either schematic form or a high level language, and then the design is automatically "compiled" and mapped to the PLD geometry, and functional and timing simulations can be run. If the simulation results are acceptable, an actual PLD can then be programmed directly, as a further step in the development process, and even tested, to some extent, with the same set of test data as was used for the simulation step. This "rapid prototyping" [19] for the production of a "chip" is not very different from the production of a working software program (and the PLD can be reprogrammed if different functionality is later desired). Such a system, of course, places many constraints on achievable designs. In addition, the automated steps, which rely on heuristics rather than exact techniques to find acceptable solutions to the many computationally complex problems that need to be solved during the development process, sacrifice performance for ease of development, so that a device designed in such a system will never achieve the ultimate performance possible for the given technology. However, the trade-offs include ease of use, much shorter development times, and the management of much larger numbers of individual circuit elements than would be possible if each individual element were tuned to its optimum performance. In addition, if a high-level language is used for input, an acceptable design can often be translated, with few changes, to a more powerful design system that will allow implementation in more flexible technologies and additional fine tuning of circuit performance. In Fig. 13.4 we see some of the levels of abstraction which are present in such a development

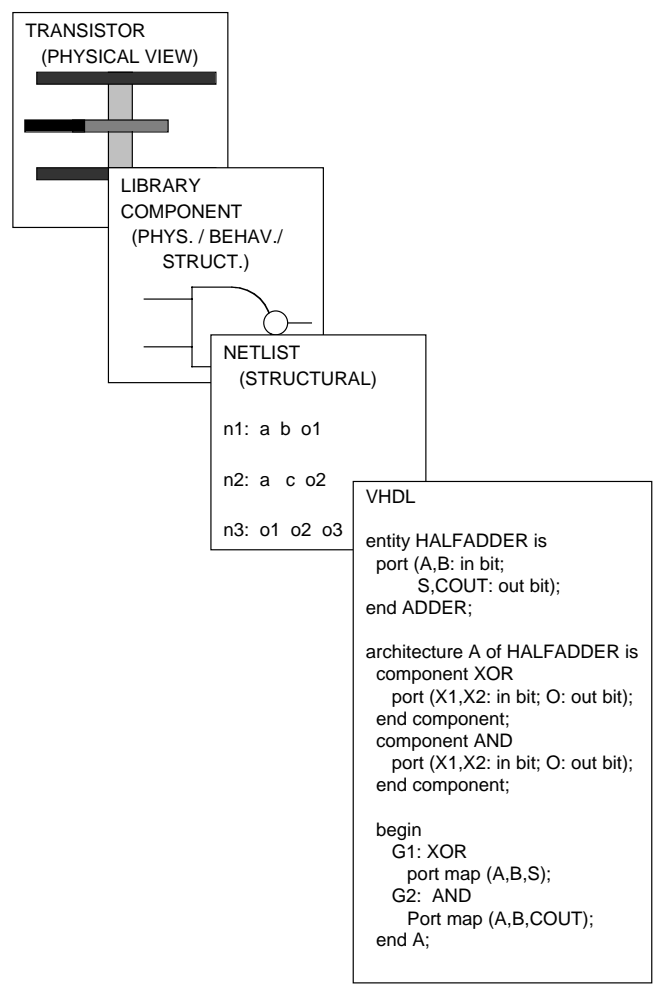



FIGURE 13.4 Levels of abstraction–half adder.

process, with the lowest level being detailed transistor models and the highest a VHDL description of a half adder.

13.3 Analog and Mixed-Signal Circuit Development: Modeling and Simulating Systems with Micro- (or Nano-) Scale Feature Sizes and Mixed Digital (Discrete) and Analog (Continuous) Input, Output, and Signals

At the lowest level, digital circuits are in fact analog devices. A CMOS inverter, for example, does not "switch" instantaneously from a voltage level representing binary 0 to a voltage level representing binary 1. However, by careful design of the inverter's physical structures, it is possible to make the switching time from the range of voltage outputs which are considered to be "0" to the range considered to be "1" (or vice versa) acceptably short. In MOSFETs, for example, the two discrete signals of interest can be identified with the transistor, modeled as a switch, being "open" or "closed," and the "switching" from one state to another can be ignored except at the very lowest levels of abstraction. In much design and simulation work, the analog aspects of the digital circuit's behavior can thus be ignored. Only at the lower levels of abstraction will the analog properties of VLSI devices or the quantum effects occurring, e.g., in a MOSFET need to be explicitly taken into account, ideally by powerful automated development tools supported by detailed models. At higher levels this behavior can be encapsulated and expressed in terms of minimum and maximum switching times with respect to a given capacitive load and given voltage levels. Even in digital systems, however, as submicron feature sizes become more common, more attention must be paid to analog effects. For example, at small feature sizes, wire delay due to RC effects and crosstalk in nearby wires become more significant factors in obtaining good simulation results [20]. It is instructive to examine how simulation support for digital systems can be extended to account for these factors.

Typically, analog circuit devices are much more likely to be "hand-crafted" than digital devices. SPICE and SPICE-like simulations are commonly used to measure performance at the level of transistors, resistors, capacitors, and inductors. For example, due to the growing importance of wireless and mobile computing, a great deal of work in analog design is currently addressing the question of how to produce circuits (digital, analog, and mixed-signal) that are "low-power," and simulations for devices to be used in these circuits are typically carried out at the SPICE level. Unless a new physical technology is to be employed, the simulations will mostly rely on the commonly available models for transistors, transmission lines, etc., thus encapsulating the lowest level behaviors.

Let us examine the factors given above for the success of digital system simulation and development to see how the analog domain compares. We assume a development cycle similar to that shown in Fig. 13.1.

- Is there a small set of basic circuit elements? In the analog domain it is possible to identify sets of components, such as current mirrors, op-amps, etc. However, there is no "universal" gate or small set of gates from which all other devices can be made, as is true in the digital domain. Another complicating factor is that elementary analog circuit elements are usually defined in terms of physical performance. There is no clean notion of 0/1 behavior. Because analog signals are continuous, it is often much more difficult to untangle complex circuit behaviors and to carry out meaningful simulations where clean parameter separations give clear results. Once a preliminary analog device or circuit design has been developed, the process of using simulations to decide on exact parameter values is known as "exploring the design space." This process necessarily exhibits high computational complexity. Often heuristic methods such as simulated annealing, neural nets, or a genetic algorithm can be used to perform the necessary search efficiently [21].
- Is there a small set of well-understood technologies? In this area, the analog and mixed signal domain is similar to the digital domain. Much analog development activity focuses on a few standard and well-parameterized technologies. In general, analog devices are much more sensitive to variations in process parameters, and this must be accounted for in analog simulation.

Statistical techniques to model process variation have been included, for example, in the APLAC tool [22], which supports object-oriented design and simulation for analog circuits. Modeling and simulation methods, which incorporate probabilistic models, will become increasingly important as nanoscale devices become more common and as new technologies depending on quantum effects and biology-based computing are developed. Several current efforts, for example, are aimed at developing a "BIOSPICE" simulator, which would incorporate more stochastic system behavior [23].

- Is there a well-developed educational infrastructure and prototyping facilities? All the organizations, which support education and prototyping in the digital domain [3–7], provide similar support for analog and mixed-signal design.
- Are encapsulation and abstraction widely employed? In the past few years, a great deal of progress has been made in incorporating these concepts into analog and mixed-signal design systems. The wide availability of very powerful computers, which can perform the necessary design and simulation tasks in reasonable amounts of time, has helped to make this progress possible. In [24], for example, top-down, constraint-driven methods are described, and in [25] a rapid prototyping method for synthesizing analog and mixed signal systems, based on the tool suite VASE (VHDL-AMS Synthesis Environment), is demonstrated. These methods rely on classifications similar to those given for digital systems in Fig. 13.2(a).
- Are there well-developed models, mature tools, and integrated development systems which are widely available? In the analog domain, there is still much more to be done in these areas than in the digital domain, but prototypes do exist. In particular, the VHDL and Verilog languages have been extended to allow for analog and mixed-signal components. The VHDL extension, e.g., VHDL-AMS [14], will allow the inclusion of any algebraic or ordinary differential equation in a simulation. However, there does not exist a completely functional VHDL-AMS simulator, although a public domain version, incorporating many useful features, is available at [26] and many commercial versions are under development (e.g., [27]). Thus, at present, expanded versions of MAGIC and SPICE are still the most widely-used design and simulation tools. While there have been some attempts to develop design systems with configurable devices similar to the digital devices shown in Fig. 13.3, these have not so far been very successful. Currently, more attention is being focused on component-based development with design reuse for SOC (systems on a chip) through initiatives such as [28].

13.4 Basic Techniques and Available Tools for MEMS Modeling and Simulation

Before trying to answer the above questions for MEMS, we need to look specifically at the tools and techniques the MEMS designer has available for the modeling and simulation tasks. As pointed out in [29,30], the bottom line is, in any simulator, all models are not created equal. The developer must be very clear about what parameters are of greatest interest and then must choose the models and simulation techniques (including implementation in a tool or tools) that are most likely to give the most accurate values for those parameters in the least amount of simulation time. For example, the model used to determine static behavior may be different from the model needed for an adequate determination of dynamic behavior. Thus, it is useful to have a range of models and techniques available.

Basic Modeling and Simulation Techniques

We need to make the following choices:

- What kind of behavior are we interested in? IC simulators, for example, typically support DC operating analysis, DC sweep analysis (stepping current or voltage source values) and transient sweep analysis (stepping time values), along with several other types of transient analysis [30].

- Will the computation be symbolic or numeric?
- Will use of an exact equation, nodal analysis, or finite element analysis be most appropriate? Currently, these are the techniques which are favored by most MEMS developers.

To show what these choices entail, let us look at a simple example that combines electrical and mechanical parts. The cantilever beam in Fig. 13.5(a), fabricated in metal, polysilicon, or a combination, may be combined with an electrically isolated plate to form a parallel plate capacitor. If a mechanical force or a varying voltage is applied to the beam (Fig. 13.5(b1)), an accelerometer or a switch can be obtained [31]. If instead the plate can be moved back and forth, a more efficient accelerometer design results (Fig. 13.5(b2)); this is the basic design of Analog Devices' accelerometer, probably the first truly successful commercial MEMS device [32,33]. If several beams are combined into two "combs," a comb-drive sensor or actuator results, as in Fig. 13.5(b3) [34]. Let us consider just the simplest case, as shown in Fig. 13.5(b1).

If we assume the force on the beam is concentrated at its end point, then we can use the method of [35] to calculate the "pull-in" voltage, i.e., the voltage at which the plates are brought together, or to a stopper which keeps the two plates from touching. We model the beam as a dampened spring-mass system and look for the force F, which, when translated into voltage, will give the correct x value for the beam to be "pulled in."

$$F = mx'' + Bx' + kx$$

Here mass $m = \rho WTL$, where ρ is the density of the beam material, $I = WT^3/12$ is the moment of inertia, $k = 3EI/L^3$, E is the Young's modulus of the beam material, and $B = (k/EI)^{1/4}$. This second-order linear differential equation can be solved numerically to obtain the pull-down voltage. In this case, since a closed form expression can be obtained for x, symbolic computation would also be an option. In [36] it is shown that for this simple problem several commonly used methods and tools will give the same result, as is to be expected.

To obtain a more accurate model of the beam we can use the method of nodal analysis, that treats the beam as a graph consisting of a set of edges or "devices," linked together at "nodes." Nodal analysis assumes that at equilibrium the sum of all values around each closed loop (the "across" quantities) will

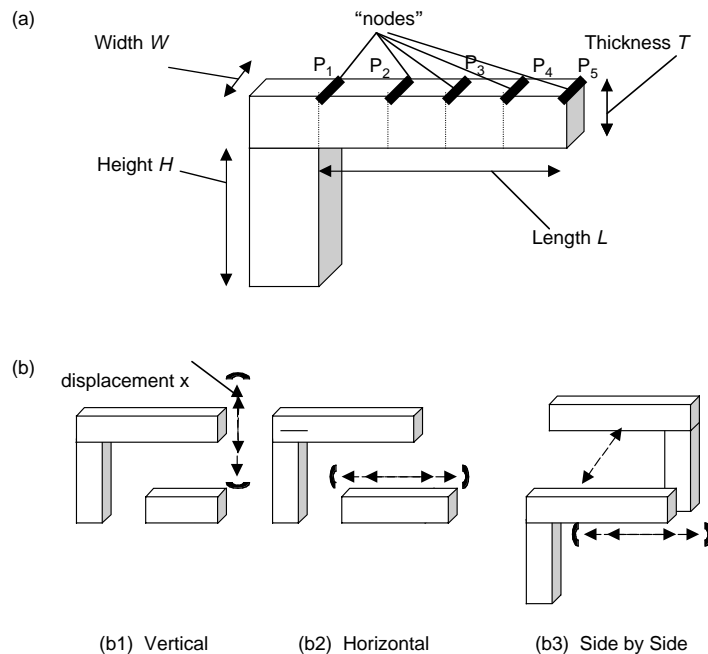



FIGURE 13.5 Cantilever beam and beam–capacitor options: (a) cantilever beam dimensions, (b) basic beam–capacitor designs.

be zero, as will the sum of all values entering or leaving a given node (the "through" quantities). Thus, for example, the sum of all forces and moments on each node must be zero, as must the sum of all currents flowing into or out of a given node. This type of modeling is sometimes referred to as "lumped parameter," since quantities such as resistance and capacitance, which are in fact distributed along a graph edge, are modeled as discrete components. In the electrical domain Kirchhoff's laws are examples of these rules. This method, which is routinely applied to electrical circuits in elementary network analysis courses (see, e.g., [37]), can easily be applied to other energy domains by using correct domain equivalents (see, e.g., [38]). A comprehensive discussion of the theory of nodal analysis can be found in [39]. In Fig. 13.5(a), the cantilever beam has been divided into four "devices," subbeams between node i and $i + 1$, $i = 1, 2, 3, 4$, where the positions of nodes i and $i + 1$ are described by (x_i, y_i, θ_i) and $(x_{i+1}, y_{i+1}, \theta_{i+1})$ the coordinates and slope at P_i and P_{i+1}. The beam is assumed to have uniform width W and thickness T, and each subbeam is treated as a two-dimensional structure free to move in three-space. In [40] a modified version of nodal analysis is used to develop numerical routines to simulate several MEMS behaviors, including static and transient behavior of a beam-capacitor actuator. This modified method also adds position coordinates z_i and z_{i+1} and replaces the slope θ_i at each node with a vector of slopes, θ_{ix}, θ_{iy}, and θ_{iz}, giving each node six degrees of freedom.

Since nodal analysis is based on linear elements represented as the edges in the underlying graph, it cannot be used to model many complex structures and phenomena such as fluid flow or piezoelectricity. Even for the cantilever beam, if the beam is composed of layers of two different materials (e.g., polysilicon and metal), it cannot be adequately modeled using nodal analysis. The technique of finite element analysis (FEA) must be used instead. For example, in some follow-up work to that reported in [36], nodal analysis and symbolic computation gave essentially the same results, but the FEA results were significantly different. Finite element analysis for the beam begins with the identification of subelements, as in Fig. 13.5(a), but each element is treated as a true three-dimensional object. Elements need not all have the same shape, for example, tetrahedral and cubic "brick" elements could be mixed together, as appropriate. In FEA, one cubic element now has eight nodes, rather than two (Fig. 13.6), so computational complexity is increased. Thus, developing efficient computer software to carry out FEA for a given structure can be a difficult task in itself. But this general method can take into account many features that cannot be adequately addressed using nodal analysis, including, for example, unaligned beam sections, and surface texture (Fig. 13.7). FEA, which can incorporate static, transient, and dynamic behavior, and which can treat heat and fluid flow, as well as electrical, mechanical, and other forces, is explained in detail in [41]. The basic procedure is as follows:

- Discretize the structure or region of interest into finite elements. These need not be homogeneous, either in size or in shape. Each element, however, should be chosen so that no sharp changes in geometry or behavior occur at an interior point.
- For each element, determine the element characteristics using a "local" coordinate system. This will represent the equilibrium state (or an approximation if that state cannot be computed exactly) for the element.
- Transform the local coordinates to a global coordinate system and "assemble" the element equations into one (matrix) equation.
- Impose any constraints implied by restricted degrees of freedom (e.g., a fixed node in a mechanical problem).
- Solve (usually numerically) for the nodal unknowns.
- From the global solution, calculate the element resultants.

A Catalog of Resources for MEMS Modeling and Simulation

To make our discussion of the state-of-the-art of MEMS simulation less confusing, we first list some of the tools and products available. This list is by no means comprehensive, but it will provide us with a range of approaches for comparison. It should be noted that this list is accurate as of July 2001, but the MEMS development community is itself developing, with both commercial companies and university

(a) Nodal analysis/Modified nodal analysis
("Linear" elements)

(b) Finite element analysis
(Three-dimensional elements)

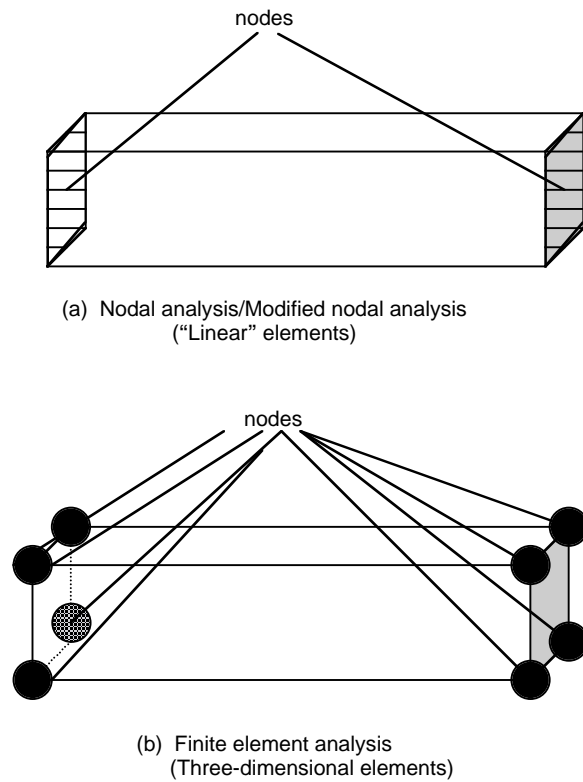



FIGURE 13.6 Nodal analysis and finite element analysis.

(a) Ideal beam

rough surface
unaligned sections

(b) Actual beam

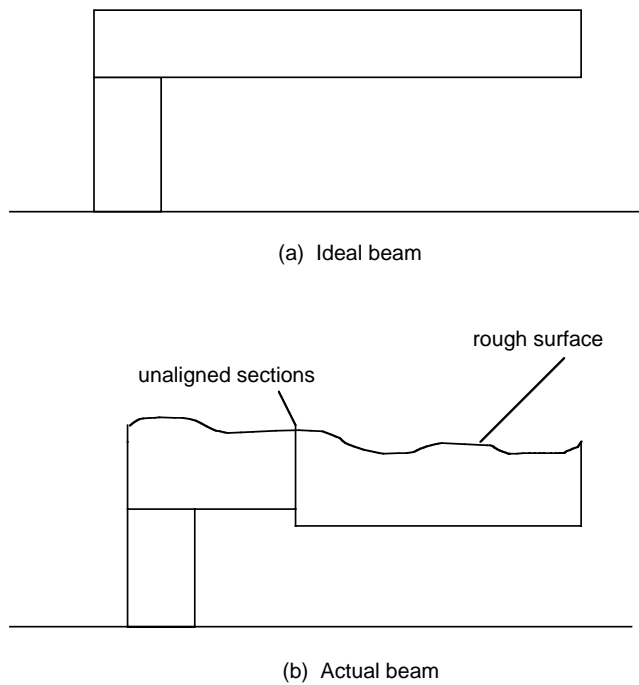



FIGURE 13.7 Ideal and actual cantilever beams (side view).

research sites frequently taking on new identities and partners and also expanding the range of services they offer.

A. Widely Available Tools for General Numeric and Symbolic Computation

These tools are relatively easy to learn to use. Most engineering students will have mastered at least one before obtaining a bachelor's degree. They can be used to model a device "from scratch" and to perform simple simulations. For more complex simulations, they are probably not appropriate for two reasons. First, neither is optimized to execute long computations efficiently. Second, developing the routines necessary to carry out a complex nodal or finite element analysis will in itself be a time-consuming task and will in most cases only replicate functionality already available in other tools listed here.

- Mathematica [42]. In [36] Mathematica simulation results for a cantilever beam-capacitor system are compared with results from several other tools.
- Matlab (integrated with Maple) [43]. In [44], for example, Matlab simulations are shown to give good approximations for a variety of parameters for microfluidic system components.

B. Tools Originally Developed for Specific Energy Domains

Low-cost easy to use versions of some of these tools (e.g., SPICE, ANSYS) are also readily available. Phenomena from other energy domains can be modeled using domain translation.

- SPICE (analog circuits) [13]. SPICE is the de facto standard for analog circuit simulators. It is also used to support simulation of transistors and other components for digital systems. SPICE implements numerical methods for nodal analysis. Several authors have used SPICE to simulate MEMS behavior in other energy domains. In [35], for example, the equation for the motion of a damped spring, which is being used to calculate pull-in voltage, is translated into the electrical domain and reasonable simulation accuracy is obtained. In [45] steady-state thermal behavior for flow-rate sensors is simulated by dividing the device to be modeled into three-dimensional "bricks," modeling each brick as a set of thermal resistors, and translating the resulting conduction and convection equations into electrical equivalents.
- APLAC [22]. This object-oriented analog and mixed-signal simulator incorporates routines, which allow statistical modeling of process variation.
- VHDL-AMS [14,26,27]. The VHDL-AMS language, designed to support digital, analog, and mixed-signal simulation, will in fact support simulation of general algebraic and ordinary differential equations. Thus mixed-energy domain simulations can be carried out. VHDL-AMS, which is typically built on a SPICE kernel, uses the technique of nodal analysis. Some VHDL-AMS MEMS models have been developed (see, e.g., [46,47]). Additional information about VHDL-AMS is available at [48].
- ANSYS [49]. Student versions of the basic ANSYS software are widely available. ANSYS is now partnering with MemsPro (see below). ANSYS models both mechanical and fluidic phenomena using FEA techniques. A survey of the ANSYS MEMS initiative can be found at [50].
- CFD software [51]. This package, which also uses FEA, was developed to model fluid flow and temperature phenomena.

C. Tools Developed Specifically for MEMS

The tools in this category use various simplifying techniques to provide reasonably accurate MEMS simulations without all the computational overhead of FEA.

- SUGAR [40,52]. This free package is built on a Matlab core. It uses nodal analysis and modified nodal analysis to model electrical and mechanical elements. Mechanical elements must be built from a fixed set of components including beams and gaps.

- NODAS v 1.4 [53]. This downloadable tool provides a library of parameterized components (beams, plate masses, anchors, vertical and horizontal electrostatic comb drives, and horizontal electrostatic gaps) that can be interconnected to form MEMS systems. The tool outputs parameters that can be used to perform electromechanical simulations with the Saber simulator [27]. A detailed example is available at [54], and a description of how the tool works (for v 1.3) is also available [55]. Useful information is also available in [70].

D. "Metatools" Which Attempt to Integrate Two or More Domain-Specific Tools into One Package

- MEMCAD, currently being supported by the firm Coventor [56]. This product was previously supported by Microcosm, Inc. It provides low-level simulation capability by integrating domain-specific FEA tools into one package to support coupled energy domain simulations. It also supports process simulation. Much of the extensive research underlying this tool is summarized in [57].
- MemsPro [58], which currently incorporates links to ANSYS. MemsPro itself is an offshoot of Tanner Tools, Inc. [59], which originally produced a version of MAGIC [12] that would run on PCs. The MemsPro system provides integrated design and simulation capability. Process "design rules" can be defined by the user. SPICE simulation capability is integrated into the toolset, and a data file for use with ANSYS can also be generated. MemsPro does not do true energy domain coupling at this time. Some library components are also available.

E. Other Useful Resources

- The MEMS Clearinghouse website [60]. This website contains links to products, research groups, and conference information. One useful link is the Material Properties database [61], which includes results from a wide number of experiments by many different research groups. Information from this database can be used for initial "back of the envelope" calculations for component feasibility, for example.
- The Cronos website [62]. This company provides prototyping and production-level fabrication for all three process approaches (surface micromachining, bulk micromachining, and high aspect ratio manufacturing). It is also attempting to build a library of MEMS components for both surface micromachining (MUMPS, or the Multi-User MEMS Process [63]) and bulk micromachining.

13.5 Modeling and Simulating MEMS, i.e., Systems with Micro- (or Nano-) Scale Feature Sizes, Mixed Digital (Discrete) and Analog (Continuous) Input, Output, and Signals, Two- and Three-Dimensional Phenomena, and Inclusion and Interaction of Multiple Domains and Technologies

In preceding sections we briefly described the current state-of-the-art in modeling and simulation in both the digital and analog domains. While the digital tools are much more developed, in both the digital and analog domains there exist standard, well-characterized technologies, standard widely available tools, and stable educational and prototyping programs. In the much more complex realm of MEMS, this is not the case. Let us compare MEMS, point by point, with digital and analog circuits.

- Is there a small set of basic elements? The answer to this question is emphatically no. Various attempts have been made by researchers to develop a comprehensive basic set of building blocks, beginning with Petersen's identification of the fundamental component set consisting of beams, membranes, holes, grooves, and joints [64]. Most of these efforts focus on adding mechanical and electromechanical elements. In the SUGAR system, for example, the basic elements are the *beam* and the *electrostatic gap*. In the Carnegie Mellon tool MEMSYN [65], which is supported by the

NODAS simulator, basic elements include beams and gaps, as well as plate masses, anchors, and electrostatic comb drives (vertical and horizontal). For the MUMPS process there is the Consolidated Micromechanical Element Library (CaMEL), which contains both a nonparameterized cell database and a library of parameterized elements (which can be accessed through a component "generator," but not directly by the user). CaMEL supports the creation of a limited set of components, including motors and resonators, in a fixed surface-micromachined technology. But the bottom line for MEMS is that no set of basic building blocks has yet been identified which can support all the designs, in many different energy domains and in a variety of technologies, which researchers are interested in building. Moreover, there is no consensus as to how to effectively limit design options so that such a fundamental set could be identified. In addition, the continuous nature of most MEMS behavior presents the same kinds of difficulties that are faced with analog elements. Development of higher level component libraries, however, is a fairly active field, with, for example, ANSYS, CFD, MEMCAD, Carnegie Mellon, and MemsPro all providing libraries of previously designed and tested components for systems developers to use. Most of these components are in the electromechanical domain. As mentioned above, a few VHDL-AMS models are also available, but these will not be of practical value until more robust and complete VHDL-AMS simulators are developed and more experimental results can be obtained to validate these models.

- Is there a small set of well-understood technologies? Again the answer must be no. Almost all digital and analog circuits are essentially two-dimensional, but, in the case of MEMS, many designs can be developed either in the "2.5-dimensional" technology known as micromachining or in the true three-dimensional technology known as bulk micromachining. Thus, before doing any modeling or simulation, the MEMS developer must first choose not only among very different fabrication techniques but also among actual processes. Both the Carnegie Mellon and Cronos tools, for example, are based on processes that are being developed in parallel with the tools. MOSIS does provide central access to technology in which all but the final steps of surface micromachining can be done, but no other centrally maintained processing is available to the community of MEMS researchers in general. For surface micromachining, the fact that the final processing steps are performed in individual research labs is problematic for producing repeatable experimental results. For bulk micromachining examples, fabrication in small research labs rather than in a production environment is more the norm than the exception, so standardization for bulk processes is difficult to achieve. In addition, because much MEMS work is relatively low-volume, most processes are not well enough characterized for low-level modeling to be very effective. In such circumstances it is very difficult to have reliable process characterizations on which to build robust models.
- Is there a well-developed educational infrastructure and prototyping facilities? Again we must answer no. Introductory MEMS courses, especially, are much more likely to emphasize fabrication techniques than modeling and simulation. In [66] a set of teaching modules for a MEMS course emphasizing integrated design and simulation is described. However, this course requires the use of devices previously fabricated for validating design and simulation results, rather than expecting students to complete the entire design-simulate-test-fabricate sequence in one quarter or semester. In addition, well-established institutional practices make it difficult to provide the necessary support for multidisciplinary education which MEMS requires.
- Are encapsulation and abstraction widely employed? In the 1980s many researchers believed that multiple levels of abstraction were not useful for MEMS devices. Currently, however, the concept of intermediate-level "macromodels" has gained much support [57,70], and increasing emphasis is being placed on developing macromodels for MEMS components that will be a part of larger systems. In addition, there are several systems in development that are based on sets of more primitive components. But this method of development is not the norm, in large part because of the rich set of possibilities inherent in MEMS in general. In Fig. 13.2(b) we have given a partial classification of MEMS corresponding to the classification for digital devices in Fig. 13.2(a). At this point it is not

Simulation Tool	Levels Supported
Mathematica, Matlab	all
MEMCAD	low
SPICE	low to medium
APLAC	low to medium
ANSYS, CFD	low to medium
SUGAR, NODAS	low to medium
MemsPro	low to medium
VHDL-AMS	medium to high

*Because MEMCAD incorporates process simulations, it supports both physical and behavioral views. All other tools support the behavioral view.

FIGURE 13.8 Available MEMS simulation tools, by level and view.

clear what the optimum number of levels of abstraction for MEMS would be. In Fig. 13.8 we have attempted to classify some of the tools from Section 13.4 in terms of their ability to support various levels (since these are simulators, they all support the "behavioral" view. MEMCAD, which allows fabrication process simulation, also supports the "physical" view). Note that VHDL-AMS is the only tool, besides the general-purpose Mathematica and Matlab, that supports a high-level view of MEMS.

- Are there well-developed models, mature tools, and integrated development systems which are widely available? While such systems do not currently exist, it is predicted that some examples should become available within the next ten years [57].

13.6 A "Recipe" for Successful MEMS Simulation

A useful set of guidelines for analog simulation can be found in [67]. From this we can construct a set of guidelines for MEMS simulation.

1. Be sure you have access to the necessary domain-specific knowledge for all energy domains of interest before undertaking the project.
2. Never use a simulator unless you know the range of answers beforehand.
3. Never simulate more of the system than is necessary.
4. Always use the simplest model that will do the job.
5. Use the simulator exactly as you would do the experiment.
6. Use a specified procedure for exploring the design space. In most cases this means that you should change only one parameter at a time.
7. Understand the simulator you are using and all the options it makes available.
8. Use the correct multipliers for all quantities.
9. Use common sense.
10. Compare your results with experiments and make them available to the MEMS community.
11. Be sensitive to the possibility of microlevel phenomena, which may make your results invalid.

The last point is particularly important. Many phenomena, which can be ignored at larger feature sizes, will need to be taken into account at the micro level. For example, at the micro scale, fluid flow can behave in dramatically different ways [44]. Many other effects of scaling feature sizes down to the microlevel, including an analysis of why horizontal cantilever beam actuators are "better" than vertical cantilever beam actuators, are discussed in Chapter 9 of [68]. Chapters 4 and 5 of [68] also provide important information for low-level modeling and simulation.

13.7 Conclusion: Continuing Progress in MEMS Modeling and Simulation

In the past fifteen years, much progress has been made in providing MEMS designers with simulators and other tools which will give them the ability to make MEMS as useful and ubiquitous as was predicted in [64]. While there is still much to be done, the future is bright for this flexible and powerful technology. One of the main challenges remaining for modeling and simulation is to complete the design and development of a high-level MEMS description language, along with supporting models and simulators, both to speed prototyping and to provide a common user-friendly language for designers. One candidate for such a language is VHDL-AMS. In [69], the strengths and weaknesses of VHDL-AMS as a tool for MEMS development are discussed. Strengths include the ability to handle both discrete and continuous behavior, smooth transitions between levels of abstraction, the ability to handle both conservative and nonconservative systems simultaneously, and the ability to import code from other languages. Major drawbacks include the inability to do symbolic computation, the limitation to ordinary differential equations, lack of support for frequency domain simulations, and inability to do automatic unit conversions. It remains to be seen whether VHDL-AMS will eventually be extended to make it more suitable to support the MEMS domain. But it is highly likely that VHDL-AMS or some similar language will eventually come to be widely used and appreciated in the MEMS community.

References

1. Kielkowski, R.M., *SPICE: Practical Device Modeling,* McGraw-Hill, 1995.
2. Leong, S.K., Extracting MOSFET RF SPICE models, http://www.polyfet.com/MTT98.pdf (accessed July 20, 2001).
3. http://www.mosis.edu (accessed July 20, 2001).
4. http://cmp.imag.fr (accessed July 20, 2001).
5. http://www.imec.be/europractice/europractice.html (accessed July 20, 2001).
6. http://www.vdec.u-tokyo.ac.jp/English (accessed July 20, 2001).
7. http://www.cmc.ca (accessed July 20, 2001).
8. Mead, C. and Conway, L., *Introduction to VLSI Systems,* Addison-Wesley, 1980.
9. Gajski, D. and Thomas, D., Introduction to silicon compilation, in *Silicon Compilation,* D. Gajski, Ed., Addison-Wesley, 1988, 1–48.
10. Weste, N. and Esraghian, K., *Principles of CMOS VLSI Design: A Systems Perspective,* 2nd ed., Addison-Wesley, 1993.
11. Foty, D., *MOSFET Modeling with SPICE,* Prentice Hall, 1997.
12. http://www.research.compaq.com/wrl/projects/magic/magic.html (accessed July 20, 2001).
13. http://bwrc.eecs.berkeley.edu/Classes/IcBook/SPICE (accessed July 20, 2001).
14. Design Automation Standards Committee, IEEE Computer Society, *IEEE VHDL Standard Language Reference Manual (Integrated with VHDL-AMS Changes),* Standard 1076.1, IEEE, 1997.
15. Ashenden, P., *The Designer's Guide to VHDL, 2nd ed.,* Morgan Kauffman, 2001.
16. Bhasker, J., *A Verilog HDL Primer,* 2nd ed., Star Galaxy Pub., 1999.
17. http://www.altera.com (accessed July 20, 2001).
18. http://www.xilinx.com (accessed July 20, 2001).
19. Hamblen, J.O. and Furman, M.D., *Rapid Prototyping of Digital Systems, A Tutorial Approach,* Kluwer, 1999.
20. Uyemura, J.P., *Introduction to VLSI Circuits and Systems,* John Wiley & Sons, Inc., 2002.
21. Sobecks, B., Performance Modeling of Analog Circuits via Neural Networks: The Design Process View, Ph.D. Dissertation, University of Cincinnati, 1998.
22. http://www.aplac.hut.fi (accessed July 20, 2001).
23. Weiss, R., Homsy, G., and Knight, T., Toward *in vivo* digital circuits, http://www.swiss.ai.mit.edu/~rweiss/bio-programming/dimacs99-evocomp-talk/ (accessed July 20, 2001).

24.Chang, H., Charbon, E., Choudhury, U., Demir, A., Liu, Felt E., Malavasi, E., Sangiovanni-Vincentelli, A., Charbon, E., and Vassiliou, I., *A Top-down, Constraint-Driven Design Methodology for Analog Integrated Circuits,* Kluwer Academic Publishers, 1996.
25.Ganesan, S., Synthesis and Rapid Prototyping of Analog and Mixed Signal Systems, Ph.D. Dissertation, University of Cincinnati, 2001.
26.SEAMS simulator project, University of Cincinnati ECECS Department, Distributed Processing Laboratory, http://www.ececs.uc.edu/~hcarter (accessed July 20, 2001).
27.http://www.analogy.com/products/Simulation/simulation.htm#Saber (accessed July 20, 2001).
28.www.design-reuse.com (accessed July 20, 2001).
29.S. M. Sandler and Analytical Engineering Inc., *The SPICE Handbook of 50 Basic Circuits,* http://dacafe.ibsystems.com/DACafe/EDATools/EDAbooks/SpiceHandBook (accessed July 20, 2001).
30.Kielkowski, R.M., *Inside Spice,* 2nd ed., McGraw Hill, 1998.
31.Gibson, D., Hare, A., Beyette, F., Jr., and Purdy, C., Design automation of MEMS systems using behavioral modeling, *Proc. Ninth Great Lakes Symposium on VLSI,* Ann Arbor Mich. (Eds. R.J. Lomax and P. Mazumder), March 1999, pp. 266–269.
32.http://www.analog.com/industry/iMEMS (accessed July 20, 2001).
33.http://www-ccrma.stanford.edu/CCRMA/Courses/252/sensors/node6.html (accessed July 20, 2001).
34.Tang, W., Electrostatic Comb Drive for Resonant Sensor and Actuator Applications, Ph.D. Dissertation, UC Berkeley, 1990.
35.Lo, N.R., Berg, E.C., Quakkelaar, S.R., Simon, J.N., Tachiki, M., Lee, H.-J., and Pister, S.J., Parameterized layout synthesis, extraction, and SPICE simulation for MEMS, *ISCAS 96,* May 1996, pp. 481–484.
36.Gibson, D., and Purdy, C.N., Extracting behavioral data from physical descriptions of MEMS for simulation, *Analog Integrated Circuits and Signal Processing 20,* 1999, pp. 227–238.
37.Hayt, W.H., Jr. and Kemmerly, J.E., *Engineering Circuit Analysis,* 5th ed., McGraw-Hill, 1993, pp. 88–95.
38.Dewey, A., Hanna, J., Hillman, B., Dussault, H., Fedder, G., Christen, E., Bakalar, K., Carter, H., and Romanowica, B., VHDL-AMS Modeling Considerations and Styles for Composite Systems, Version 2.0, http://www.ee.duke.edu/research/IMPACT/documents/model_g.pdf (accessed July 20, 2001).
39.McCalla, W.J., *Fundamentals of Computer-Aided Circuit Simulation,* Kluwer Academic, 1988.
40.Clark, J.V., Zhou, N., and Pister, K.S.J., Modified nodal analysis for MEMS with multi-energy domains, *International Conference on Modeling and Simulation of Microsystems, Semiconductors, Sensors and Actuators,* San Diego, CA, March 27–29, 2000, pp. 31–34.
41.Stasa, F.L., *Applied Finite Element Analysis for Engineers,* Holt, Rinehart and Winston, 1985.
42.http://www.wolfram.com/products/mathematica (accessed July 20, 2001).
43.http://www.mathworks.com/products/matlab (accessed July 20, 2001).
44.Mehta, A., Design and Control Oriented Approach to the Modeling of Microfluidic System Components, M.S. Thesis, University of Cincinnati, 1999.
45.Swart, N., Nathan, A., Shams, M., and Parameswaran, M., Numerical optimisation of flow-rate microsensors using circuit simulation tools, *Transducers '91,* 1991, pp. 26–29.
46.http://www.ee.duke.edu/research/IMPACT/vhdl-ams/index.html (accessed July 20, 2001).
47.Gibson, D., Carter, H., and Purdy, C., The use of hardware description languages in the development of microelectromechanical systems, *International Journal of Analog Integrated Circuits and Signal Processing,* 28(2), August 2001, pp. 173–180.
48.http://www.vhdl-ams.com/ (accessed July 20, 2001).
49.http://www.ansys.com/action/MEMSinitiative/index.htm (accessed July 20, 2001).
50.http://www.ansys.com/action/pdf/MEMS_WP.pdf (accessed July 20, 2001).
51.http://www.cfdrc.com (accessed July 20, 2001).
52.Pister, K., SUGAR V2.0, http://www-bsac.EECS.Berkeley.edu/~cfm/ mainpage.html (accessed July 20, 2001).
53.http://www.ece.cmu.edu/~mems/projects/memsyn/nodasv1_4/index.shtml (accessed July 20, 2001).
54.http://www2.ece.cmu.edu/~mems/projects/memsyn/nodasv1_4/tutorial.html (accessed July 20, 2001).

55. Jing, Q. and Fedder, G.K., NODAS 1.3-nodal design of actuators and sensors, *IEEE/VIUF International Workshop on Bahavioral Modeling and Simulation,* Orlando, Fla., October 27–28, 1998.
56. http://www.coventor.com/software/coventorware/index.html (accessed July 20, 2001).
57. Senturia, S.D., Simulation and design of microsystems: a 10-year perspective, *Sensors and Actuators A,* 67, 1998, pp. 1–7.
58. www.memscap.com/index2.html (accessed July 20, 2001).
59. http://www.tanner.com/ (accessed July 20, 2001).
60. http://mems.isi.edu (accessed July 20, 2001).
61. http://mems.isi.edu/mems/materials/index.html (accessed July 20, 2001).
62. http://www.memsrus.com (accessed July 20, 2001).
63. http://www.memsrus.com/cronos/svcsmumps.html (accessed July 20, 2001).
64. Petersen, K., Silicon as a mechanical material, *IEEE Proceedings,* 70(5), May 1982, pp. 420–457.
65. http://www.ece.cmu.edu/~mems/projects/memsyn/index.shtml (accessed July 20, 2001).
66. Beyette, F., Jr. and C.N. Purdy, Teaching modules for a class in mechatronics, *European Workshop on Microelectronics Education (EWME2000),* May 2000.
67. Allen, P.E. and Holberg, D.R., *CMOS Analog Circuit Design,* Oxford University Press, 1987, pp. 142–144.
68. Madou, M., *Fundamentals of Microfabrication,* CRC Press, Boca Raton, FL, 1997.
69. Gibson, D. and Purdy, C., The strengths and weaknesses of VHDL-AMS as a tool for MEMS development, white paper, 2000, http://www.ececs.uc.edu/~cpurdy/csl.html/pub.html/weakvhdl. pdf (accessed July 20, 2001).
70. Mukherjee, T. and Fedder, G.K., Hierarchical mixed-domain circuit simulation, synthesis and extraction methodology for MEMS, *Journal of VLSI Signal Processing,* 21, 1999, pp. 233–249.

14

Rotational and Translational Microelectromechanical Systems: MEMS Synthesis, Microfabrication, Analysis, and Optimization

Sergey Edward Lyshevski
Purdue University Indianapolis

14.1 Introduction

Electromagnetic-based MEMS are widely used in various sensing and actuation applications. For these MEMS, rotational and translational motion microdevices are needed to be devised, designed, and controlled. We introduce the classifier paradigm to perform the structural synthesis of MEMS upon electromagnetic features. As motion microdevices are devised, the following issues are emphasized: modeling, analysis, simulation, control, optimization, and validation. Innovative results are researched and studied applying the classifier, structural synthesis, design, analysis, and optimization concepts developed. The need for innovative integrated methods to perform the comprehensive analysis, high-fidelity modeling, and design of MEMS has facilitated theoretical developments within the overall spectrum of engineering and science. This chapter provides one with viable tools to perform structural synthesis, modeling, analysis, optimization, and control of MEMS.

Microelectromechanical systems integrate motion microstructures and devices as well as ICs on a single chip or on a hybrid chip. To fabricate MEMS, modified advanced microelectronics fabrication technologies, techniques, processes, and materials are used. Due to the use of complementary metal oxide semiconductor (CMOS) lithography-based technologies in fabrication microstructures, microdevices, and ICs, MEMS leverage microelectronics.

The following definition for MEMS was given in [1]:

> Batch-fabricated microscale devices (ICs and motion microstructures) that convert physical parameters to electrical signals and vice versa, and in addition, microscale features of mechanical and electrical components, architectures, structures, and parameters are important elements of their operation and design.

The scope of MEMS has been further expanded towards devising novel paradigms, system-level integration high-fidelity modeling, data-intensive analysis, control, optimization, fabrication, and implementation. Therefore, we define MEMS as:

Batch-fabricated microscale systems (motion and radiating energy microdevices/microstructures—driving/sensing circuitry—controlling/processing ICs) that

1. convert physical stimuli, events, and parameters to electrical and mechanical signals and vice versa,
2. perform actuation and sensing,
3. comprise control (intelligence, decision making, evolutionary learning, adaptation, self-organization, etc.), diagnostics, signal processing, and data acquisition features,

and microscale features of electromechanical, electronic, optical, and biological components (structures, devices, and subsystems), architectures, and operating principles are basics of their operation, design, analysis, and fabrication.

The integrated design, analysis, optimization, and virtual prototyping of intelligent and high-performance MEMS, system intelligence, learning, adaptation, decision making, and self-organization can be addressed, researched, and solved through the use of advanced electromechanical theory, state-of-the-art hardware, novel technologies, and leading-edge software. Many problems in MEMS can be formulated, attacked, and solved using the microelectromechanics. In particular, microelectromechanics deals with benchmarking and emerging problems in integrated electrical–mechanical–computer engineering, science, and technologies. Microelectromechanics is the integrated design, analysis, optimization, and virtual prototyping of high-performance MEMS, system intelligence, learning, adaptation, decision making, and control through the use of advanced hardware, leading-edge software, and novel fabrication technologies and processes. Integrated multidisciplinary features approach quickly, and the microelectromechanics takes place.

The computer-aided design tools are required to support MEMS analysis, simulation, design, optimization, and fabrication. Much effort has been devoted to attain the specified steady-state and dynamic performance of MEMS to meet the criteria and requirements imposed. Currently, MEMS are designed, optimized, and analyzed using available software packages based on the linear and steady-state analysis.

However, highly detailed nonlinear electromagnetic and mechanical modeling must be performed to design high-performance MEMS. Therefore, the research is concentrated on high-fidelity mathematical modeling, data intensive analysis, and nonlinear simulations, as well as control (design of control algorithms to attain the desired performance). The reported synthesis, modeling, analysis, simulation, optimization, and control concepts, tools, and paradigms ensure a cost-effective solution and can be used to guarantee rapid prototyping of high-performance state-of-the-art MEMS. It is often very difficult, and sometimes impossible, to solve a large array of nonlinear analysis and design problems for motion microdevices using conventional methods. Innovative concepts, methods, and tools that fully support the analysis, modeling, simulation, control, design, and optimization are needed. The fabrication technologies used in MEMS were developed [2,3], and micromachining technologies are discussed in this chapter. This chapter solves a number of long-standing problems for electromagnetic-based MEMS.

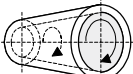
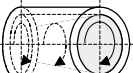
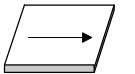
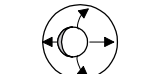
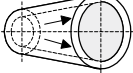
14.2 MEMS Motion Microdevice Classifier and Structural Synthesis

It was emphasized that the designer must design MEMS by devising novel high-performance motion microdevices, radiating energy microdevices, microscale driving/sensing circuitry, and controlling/processing ICs. A step-by-step procedure in the design of motion microdevices is:

- define application and environmental requirements,
- specify performance specifications,
- devise motion microstructures and microdevices, radiating energy microdevices, microscale driving/sensing circuitry, and controlling/processing ICs,
- develop the fabrication process using micromachining and CMOS technologies,
- perform electromagnetic, energy conversion, mechanical, and sizing/dimension estimates,
- perform electromagnetic, mechanical, vibroacoustic, and thermodynamic design with performance analysis and outcome prediction,
- verify, modify, and refine design with ultimate goals and objectives to optimize the performance.

In this section, the design and optimization of motion microdevices is reported.

To illustrate the procedure, consider two-phase permanent-magnet synchronous slotless micromachines as documented in Fig. 14.1.

It is evident that the electromagnetic system is *endless*, and different geometries can be utilized as shown in Fig. 14.1. In contrast, in translational (linear) synchronous micromachines, the *open-ended* electromagnetic system results. The attempts to classify microelectromechanical motion devices were made in [1,4,5]; however, the qualitative and quantitative comprehensive analysis must be researched.

Motion microstructure geometry and electromagnetic systems must be integrated into the synthesis, analysis, design, and optimization. Motion microstructures can have the plate, spherical, torroidal, conical, cylindrical, and asymmetrical geometry. Using these distinct geometry and electromagnetic systems, we propose to classify MEMS. This idea is extremely useful in the study of existing MEMS as well as in the synthesis of an infinite number of innovative motion microdevices. In particular, using the possible geometry and electromagnetic systems (*endless, open-ended*, and *integrated*), novel high-performance MEMS can be synthesized.

The basic electromagnetic micromachines (microdevices) under consideration are direct- and alternating-current, induction and synchronous, rotational and translational (linear). That is, microdevices are classified using a type classifier

$$Y = \{y : y \in \quad Y\}$$

Motion microdevices are categorized using a geometric classifier (plate P, spherical S, torroidal T, conical N, cylindrical C, or asymmetrical A geometry) and an electromagnetic system classifier (*endless* E, *open-ended* O, or *integrated* I). The microdevice classifier, documented in Table 14.1, is partitioned

TABLE 14.1 Classification of Electromagnetic Microdevices Using the Electromagnetic System–Geometry Classifier

M \ G		Geometry					
		Plate, *P*	Spherical, *S*	Torroidal, *T*	Conical, *N*	Cylindrical, *C*	Asymmetrical, *A*
Electromagnetic System	*Endless* (Closed), *E*						Σ
	Open-Ended (Open), *O*						Σ
	Integrated, I	Σ	Σ	Σ	Σ	Σ	Σ

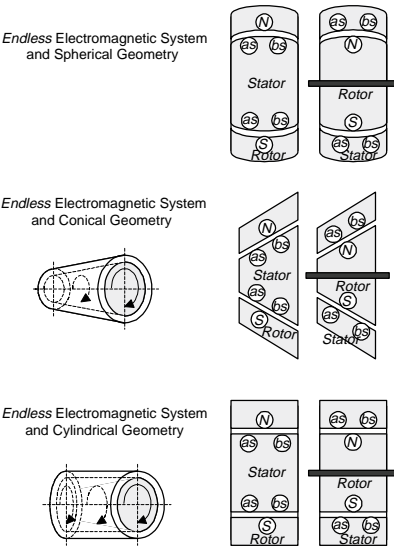



FIGURE 14.1 Permanent-magnet synchronous micromachines with different geometry.

into three horizontal and six vertical strips, and contains 18 sections, each identified by ordered pairs of characters, such as (E, P) or (O, C).

In each ordered pair, the first entry is a letter chosen from the bounded electromagnetic system set

$$M = \{E, O, I\}$$

The second entry is a letter chosen from the geometric set

$$G = \{P, S, T, N, C, A\}$$

That is, for electromagnetic microdevices, the electromagnetic system–geometric set is

$$M \times G = \{(E, F), (E, S), (E, T), \ldots, (I, N), (I, C), (I, A)\}$$

In general, we have

$$M \times G = \{(m, g) : m \in M \text{ and } g \in G\}$$

Other categorization can be applied. For example, single-, two-, three-, and multi-phase microdevices are classified using a phase classifier

$$H = \{h : h \in H\}$$

Therefore, $Y \times M \times G \times H = \{(y, m, g, h) : y \in Y, m \in M, g \in G \text{ and } h \in H\}$

Topology (radial or axial), permanent magnets shaping (strip, arc, disk, rectangular, triangular, or other shapes), permanent magnet characteristics (*BH* demagnetization curve, energy product, hysterisis minor loop), commutation, *emf* distribution, cooling, power, torque, size, torque-speed characteristics, as well as other distinct features of microdevices can be easily classified.

That is, the devised electromagnetic microdevices can be classified by an *N*-tuple as

{microdevice type, electromagnetic system, geometry, topology, phase, winding, connection, cooling}.

Using the classifier, which is given in Table 14.1 in terms of electromagnetic system–geometry, the designer can classify the existing motion microdevices as well as synthesize novel high-performance microdevices. As an example, the spherical, conical, and cylindrical geometries of a two-phase permanent-magnet synchronous microdevice are illustrated in Fig. 14.2.

This section documents new results in structural synthesis which can be used to optimize the microdevice performance. The conical (existing) and spherical-conical (devised) microdevice geometries are illustrated in Fig. 14.2. Using the innovative spherical-conical geometry, which is different compared to the existing conical geometry, one increases the active length L_r and average diameter D_r. For radial flux microdevices, the electromagnetic torque T_e is proportional to the squared rotor diameter and axial length. In particular, $T_e = k_T D_r^2 L_r$, where k_T is the constant. From the above relationship, it is evident

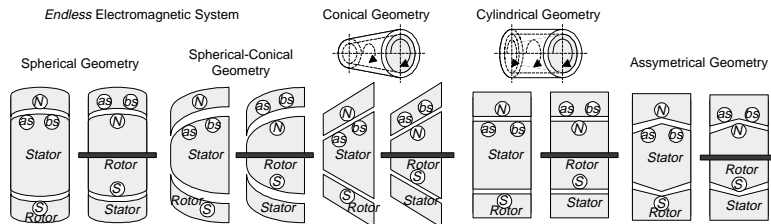



FIGURE 14.2 Two-phase permanent-magnet synchronous microdevice (micromachine) geometry.

that the spherical-conical micromotors develop higher electromagnetic torque compared with the conventional design. In addition, improved cooling, reduced undesirable torques components, as well as increased ruggedness and robustness contribute to the viability of the proposed solution. Thus, using the classifier paradigm, novel microdevices with superior performance can be devised.

14.3 MEMS Fabrication

Microelectromechanics, which integrates micromechanics and microelectronics, requires affordable, low-cost, high-yield fabrication technologies which allow one to fabricate 3-D microscale structures and devices. Micromachining is a key fabrication technology for microscale structures, devices, and MEMS. Microelectromechanical systems fabrication technologies fall into three broad categories: bulk machining, surface machining, and LIGA (LIGA-like) techniques [1–3].

Bulk Micromachining

Bulk and surface micromachining are based on the modified CMOS and specifically designed micromachining processes. Bulk micromachining of silicon uses wet and dry etching techniques in conjunction with etch masks and etch-stop-layers to develop microstructures from the silicon substrate. Microstructures are fabricated by etching areas of the silicon substrate to release the desired 3-D microstructures. The *anisotropic* and *isotropic* wet etching processes, as well as concentration dependent etching techniques, are widely used in bulk micromachining. The microstructures are formed by etching away the bulk of the silicon wafer to fabricate the desired 3-D structures. Bulk machining with its crystallographic and dopant-dependent etch processes, when combined with wafer-to-wafer bonding, produces complex 3-D microstructures with the desired geometry. Through bulk micromachining, one fabricates microstructures by etching deeply into the silicon wafer. There are several ways to etch the silicon wafer. The *anisotropic* etching uses etchants that etch different crystallographic directions at different rates. Through *anisotropic* etching, 3-D structures (cons, pyramids, cubes, and channels into the surface of the silicon wafer) are fabricated. In contrast, the *isotropic* etching etches all directions in the silicon wafer at same (or close) rate, and, therefore, hemisphere and cylinder structures can be made. Deep reactive ion etching uses plasma to etch straight walled structures (cubes, rectangular, triangular, etc.).

Surface Micromachining

Surface micromachining has become the major fabrication technology in recent years because complex 3-D microscale structures and devices can be fabricated. Surface micromachining with single-crystal silicon, polysilicon, silicon nitride, silicon oxide, and silicon dioxide (as structural and sacrificial materials which deposited and etched) is widely used to fabricate microscale structures and devices on the surface of a silicon wafer. This affordable low-cost high-yield technology is integrated with IC fabrication processes guaranteeing the needed microstructures-IC fabrication compatibility. The techniques for depositing and patterning thin films are used to produce complex microstructures and microdevices on the surface of silicon wafers (surface silicon micromachining) or on the surface of other substrates. Surface micromachining technology allows one to fabricate the structure as layers of thin films. This technology guarantees the fabrication of 3-D microdevices with high accuracy, and the surface micromachining can be called a thin film process. Each thin film is usually limited to thickness up to 5 μm, which leads to fabrication of high-performance planar-type microscale structures and devices. The advantage of surface micromachining is the use of standard CMOS fabrication processes and facilities, as well as compliance with ICs. Therefore, this technology is widely used to manufacture microscale actuators and sensors (microdevices).

Surface micromachining is based on the application of sacrificial (temporary) layers that are used to maintain subsequent layers and are removed to reveal (release) fabricated (released or suspended) microstructures. This technology was first demonstrated for ICs and applied to fabricate microstructures in the 80s. On the surface of a silicon wafer, thin layers of structural and sacrificial materials are deposited

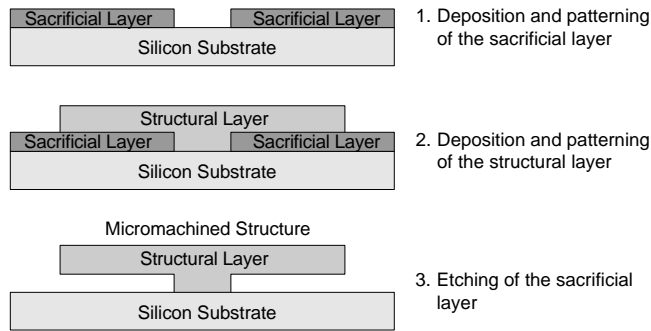



FIGURE 14.3 Surface micromachining.

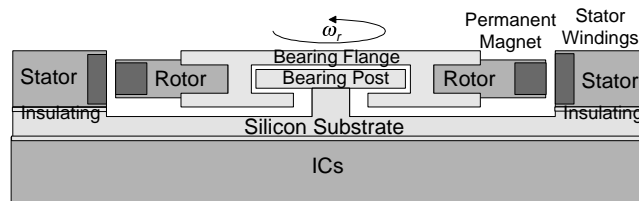



FIGURE 14.4 Cross-section schematics for slotless permanent-magnet brushless micromotor with ICs.

and patterned. Then, the sacrificial material is removed, and a micromechanical structure or device is fabricated. Figure 14.3 illustrates a typical process sequence of the surface micromachining fabrication technology.

Usually, the sacrificial layer is made of silicon dioxide (SiO_2), phosphorous-doped silicon dioxide, or silicon nitride (Si_3N_4). The structural layers are then typically formed with polysilicon, and the sacrificial layer is removed. In particular, after fabrication of the surface microstructures and microdevices (micromachines), the silicon wafer can be wet bulk etched to form cavities below the surface components, which allows a wider range of desired motion for the device. The wet etching can be done using hydrofluoric and buffered hydrofluoric acids, potassium hydroxide, ethylene-diamene-pyrocatecol, tetramethylammonium hydroxide, or sodium hydroxide. Surface micromachining technology was used to fabricate rotational micromachines [6]. For example, heavily-phosphorous-doped polysilicon can be used to fabricate rotors and stators, and silicon nitride can be applied as the structural material to attain electrical insulation. The cross-section of the slotless micromotor fabricated on the silicon substrate with polysilicon stator with deposited windings, polysilicon rotor with deposited permanent-magnets, and bearing is illustrated in Fig. 14.4. The micromotor is controlled by the driving/sensing and controlling/processing ICs. To fabricate micromotor and ICs on a single- or double-sided chip (which significantly enhances the performance), similar fabrication technologies and processes are used, and the compatibility issues are addressed and resolved. The surface micromachining processes were integrated with the CMOS technology (e.g., similar materials, lithography, etching, and other techniques). To fabricate the integrated MEMS, post-, mixed-, and pre-CMOS/micromachining techniques can be applied [1–3].

LIGA and LIGA-Like Technologies

There is a critical need to develop the fabrication technologies allowing one to fabricate high-aspect-ratio microstructures. The LIGA process, which denotes Lithography–Galvanoforming–Molding (in German words, *L*ithografie–*G*alvanik–*A*bformung), is capable of producing 3-D microstructures of up to centimeter high with the aspect ratio (depth versus lateral dimension) more than 100 [2,7,8]. The LIGA technology is based upon X-ray lithography, which guarantees shorter wavelength (in order from

few to 10 Å, which leads to negligible diffraction effects) and larger depth of focus compared with optical lithography. The ability to fabricate microstructures and microdevices in the centimeter range is particularly important in the actuators and drives applications since the specifications are imposed on the rated force and torque developed by the microdevices, and due to the limited force and torque densities, the designer faces the need to increase the actuator dimensions.

14.4 MEMS Electromagnetic Fundamentals and Modeling

The MEMS classifier, structural synthesis, and optimization were reported in Section 14.2. The classification and optimization are based on the consideration and synthesis of the electromagnetic system, analysis of the *magnetomotive* force, design of the MEMS geometry and topology, and optimization of other quantities. Different rotational (radial and axial) and translational motion microdevices are classified using *endless* (closed), *open-ended* (open), and *integrated* electromagnetic systems.

Our goal is to approach and solve a wide range of practical problems encountered in nonlinear design, modeling, analysis, control, and optimization of motion microstructures and microdevices with driving/sensing circuitry controlled by ICs for high-performance MEMS. Studying MEMS, the emphases are placed on:

- design of high-performance MEMS through devising innovative motion microdevices with radiating energy microdevices, microscale driving/sensing circuitry, and controlling/signal processing ICs,
- optimization and analysis of rotational and translation motion microdevices,
- development of high-performance signal processing and controlling ICs for microdevices devised,
- development of mathematical models with minimum level of simplifications and assumptions in the time domain,
- design of optimal robust control algorithms,
- design of intelligent systems through self-adaptation, self-organization, evolutionary learning, decision-making, and intelligence,
- development of advanced software and hardware to attain the highest degree of intelligence, integration, efficiency, and performance.

In this section, our goal is to perform nonlinear modeling, analysis, and simulation. To attain these objectives, we apply the MEMS synthesis paradigm, develop nonlinear mathematical models to model complex electromagnetic-mechanical dynamics, perform optimization, design closed-loop control systems, and perform data-intensive analysis in the time domain.

To model electromagnetic motion microdevices, using the magnetic vector and electric scalar potentials $\vec{A}$ and V, respectively, one usually solves the partial differential equations

$$-\nabla^2 \vec{A} + \mu\sigma \frac{\partial \vec{A}}{\partial t} + \mu\varepsilon \frac{\partial^2 \vec{A}}{\partial t^2} = -\mu\sigma\nabla V$$

using finite element analysis. Here, μ, σ, and ε are the permeability, conductivity, and permittivity.

However, to design electromagnetic MEMS as well as to perform electromagnetic–mechanical analysis and optimization, differential equations must be solved in the time domain. In fact, basic phenomena cannot be comprehensively modeled, analyzed, and assessed applying traditional finite element analysis, which gives the steady-state solutions and models. There is a critical need to develop the modeling tools that will allow one to augment nonlinear electromagnetics and mechanics in a single electromagnetic–mechanical modeling core to attain high-fidelity analysis with performance assessment and outcome prediction.

Operating principles of MEMS are based upon electromagnetic principles. A complete electromagnetic model is derived in terms of five electromagnetic field vectors. In particular, three electric field vectors

and two magnetic field vectors are used. The electric field vectors are the electric field intensity, $\vec{E}$, the electric flux density, $\vec{D}$, and the current density, $\vec{J}$. The magnetic field vectors are the magnetic field intensity $\vec{H}$ and the magnetic field density $\vec{B}$. The differential equations for microelectromechanical motion device are found using Maxwell's equations, constitutive (auxiliary) equations, and classical mechanics.

Maxwell's partial differential equations in the $\vec{E}$- and $\vec{H}$-domain in the point form are

$$\nabla \times \vec{E}(x, y, z, t) = -\mu \frac{\partial \vec{H}(x, y, z, t)}{\partial t}$$

$$\nabla \times \vec{H}(x, y, z, t) = \varepsilon \frac{\partial \vec{E}(x, y, z, t)}{\partial t} + \vec{J}(x, y, z, t) = \varepsilon \frac{\partial \vec{E}(x, y, z, t)}{\partial t} + \sigma \vec{E}(x, y, z, t)$$

$$\nabla \cdot \vec{E}(x, y, z, t) = \frac{\rho_v(x, y, z, t)}{\varepsilon}$$

$$\nabla \cdot \vec{H}(x, y, z, t) = 0$$

where ε is the permittivity, μ is the permeability, σ is the conductivity, and ρ_v is the volume charge density.

The constitutive (auxiliary) equations are given using the permittivity ε, permeability tensor μ, and conductivity σ. In particular, one has

$$\vec{D} = \varepsilon \vec{E} \quad \text{or} \quad \vec{D} = \varepsilon \vec{E} + \vec{P}$$

$$\vec{B} = \mu \vec{H} \quad \text{or} \quad \vec{B} = \mu(\vec{H} + \vec{M})$$

$$\vec{J} = \sigma \vec{E} \quad \text{or} \quad \vec{J} = \rho_v \vec{v}$$

The Maxwell's equations can be solved using the boundary conditions on the field vectors. In two-region media, we have

$$\vec{a}_N \times (\vec{E}_2 - \vec{E}_1) = 0, \quad \vec{a}_N \times (\vec{H}_2 - \vec{H}_1) = \vec{J}_s, \quad \vec{a}_N \cdot (\vec{D}_2 - \vec{D}_1) = \rho_s, \quad \vec{a}_N \cdot (\vec{B}_2 - \vec{B}_1) = 0$$

where $\vec{J}_s$ is the surface current density vector, $\vec{a}_N$ is the surface normal unit vector at the boundary from region 2 into region 1, and ρ_s is the surface charge density.

The constitutive relations that describe media can be integrated with Maxwell's equations, which relate the fields in order to find two partial differential equations. Using the electric and magnetic field intensities $\vec{E}$ and $\vec{H}$ to model electromagnetic fields in MEMS, one has

$$\nabla \times (\nabla \times \vec{E}) = \nabla(\nabla \cdot \vec{E}) - \nabla^2 \vec{E} = -\mu \frac{\partial \vec{J}}{\partial t} - \mu \frac{\partial^2 \vec{D}}{\partial t^2} = -\mu\sigma \frac{\partial \vec{E}}{\partial t} - \mu\varepsilon \frac{\partial^2 \vec{E}}{\partial t^2}$$

$$\nabla \times (\nabla \times \vec{H}) = \nabla(\nabla \cdot \vec{H}) - \nabla^2 \vec{H} = -\mu\sigma \frac{\partial \vec{H}}{\partial t} - \mu\varepsilon \frac{\partial^2 \vec{H}}{\partial t^2}$$

The following pair of homogeneous and inhomogeneous wave equations

$$\nabla^2 \vec{E} - \mu\sigma \frac{\partial \vec{E}}{\partial t} - \mu\varepsilon \frac{\partial^2 \vec{E}}{\partial t^2} = \nabla\left(\frac{\rho_v}{\varepsilon}\right)$$

$$\nabla^2 \vec{H} - \mu\sigma \frac{\partial \vec{H}}{\partial t} - \mu\varepsilon \frac{\partial^2 \vec{H}}{\partial t^2} = 0$$

is equivalent to four Maxwell's equations and constitutive relations. For some cases, these two equations can be solved independently. It must be emphasized that it is not always possible to use the boundary conditions using only $\vec{E}$ and $\vec{H}$, and thus, the problem not always can be simplified to two electromagnetic field vectors. Therefore, the electric scalar and magnetic vector potentials are used. Denoting the magnetic vector potential as $\vec{A}$ and the electric scalar potential as V, we have

$$\nabla \times \vec{A} = \vec{B} = \mu \vec{H} \quad \text{and} \quad \vec{E} = -\frac{\partial \vec{A}}{\partial t} - \nabla V$$

The electromagnetic field is derivative from the potentials. Using the Lorentz equation

$$\nabla \cdot \vec{A} = -\frac{\partial V}{\partial t}$$

the inhomogeneous vector potential wave equation to be solved is

$$-\nabla^2 \vec{A} + \mu\sigma \frac{\partial \vec{A}}{\partial t} + \mu\varepsilon \frac{\partial^2 \vec{A}}{\partial t^2} = -\mu\sigma\nabla V$$

To model motion microdevices, the mechanical equations must be used, and Newton's second law is usually applied to derive the equations of motion.

Using the volume charge density ρ_v, the Lorenz force, which relates the electromagnetic and mechanical phenomena, is found as

$$\vec{F} = \rho_v(\vec{E} + \vec{v} \times \vec{B}) = \rho_v \vec{E} + \vec{J} \times \vec{B}$$

The electromagnetic force can be found by applying the Maxwell stress tensor method. This concept employs a volume integral to obtain the stored energy, and stress at all points of a bounding surface can be determined. The sum of local stresses gives the net force. In particular, the electromagnetic stress is

$$\vec{F} = \int_v (\rho_v \vec{E} + \vec{J} \times \vec{B}) dv = \frac{1}{\mu} \oint_s \overleftrightarrow{T}_{\alpha\beta} \cdot d\vec{s}$$

The electromagnetic stress energy tensor (the second Maxwell stress tensor) is

$$\overleftrightarrow{T}_{\alpha\beta} = \begin{bmatrix} 0 & \vec{E}_x & \vec{E}_y & \vec{E}_z \\ -\vec{E}_x & 0 & \vec{B}_z & -\vec{B}_y \\ -\vec{E}_y & -\vec{B}_z & 0 & \vec{B}_x \\ -\vec{E}_z & \vec{B}_y & -\vec{B}_x & 0 \end{bmatrix}$$

In general, the electromagnetic torque developed by motion microstructures is found using the electromagnetic field. In particular, the electromagnetic stress tensor is given as

$$T_s = T_s^E + T_s^M$$

$$= \begin{bmatrix} E_1D_1 - \frac{1}{2}E_jD_j & E_1D_2 & E_1D_3 \\ E_2D_1 & E_2D_2 - \frac{1}{2}E_jD_j & E_2D_3 \\ E_3D_1 & E_3D_2 & E_3D_3 - \frac{1}{2}E_jD_j \end{bmatrix} + \begin{bmatrix} B_1H_1 - \frac{1}{2}B_jH_j & B_1H_2 & B_1H_3 \\ B_2H_1 & B_2H_2 - \frac{1}{2}B_jH_j & B_2H_3 \\ B_3H_1 & B_3H_2 & B_3H_3 - \frac{1}{2}B_jH_j \end{bmatrix}$$

For the Cartesian, cylindrical, and spherical coordinate systems, which can be used to develop the mathematical model, we have

$$E_x = E_1,\ E_y = E_2,\ E_z = E_3, \quad D_x = D_1,\ D_y = D_2,\ D_z = D_3,$$
$$H_x = H_1,\ H_y = H_2,\ H_z = H_3, \quad B_x = B_1,\ B_y = B_2,\ B_z = B_3$$
$$E_r = E_1,\ E_\theta = E_2,\ E_z = E_3, \quad D_r = D_1,\ D_\theta = D_2,\ D_z = D_3,$$
$$H_r = H_1,\ H_\theta = H_2,\ H_z = H_3, \quad B_r = B_1,\ B_\theta = B_2,\ B_z = B_3$$
$$E_\rho = E_1,\ E_\theta = E_2,\ E_\phi = E_3, \quad D_\rho = D_1,\ D_\theta = D_2,\ D_\phi = D_3,$$
$$H_\rho = H_1,\ H_\theta = H_2,\ H_\phi = H_3, \quad B_\rho = B_1,\ B_\theta = B_2,\ B_\phi = B_3$$

Maxwell's equations can be solved using the MATLAB environment.

In motion microdevices, the designer analyzes the torque or force production mechanisms.

Newton's second law for rotational and translational motions is

$$\frac{d\omega_r}{dt} = \frac{1}{J}\sum \vec{T}_\Sigma, \qquad \frac{d\theta_r}{dt} = \omega_r$$
$$\frac{dv}{dt} = \frac{1}{m}\sum \vec{F}_\Sigma, \qquad \frac{dx}{dt} = v$$

where ω_r and θ_r are the angular velocity and displacement, v and x are the linear velocity and displacement, $\sum \vec{T}_\Sigma$ is the net torque, $\sum \vec{F}_\Sigma$ is the net force, J is the equivalent moment of inertia, and m is the mass.

14.5 MEMS Mathematical Models

The problems of modeling and control of MEMS are very important in many applications. A mathematical model is a mathematical description (in the form of functions or equations) of MEMS, which integrate motion microdevices (microscale actuators and sensors), radiating energy microdevices, microscale driving/sensing circuitry, and controlling/signal processing ICs. The purpose of the model development is to understand and comprehend the phenomena, as well as to analyze the end-to-end behavior.

To model MEMS, advanced analysis methods are required to accurately cope with the involved highly complex physical phenomena, effects, and processes. The need for high-fidelity analysis, computationally-efficient algorithms, and simulation time reduction increases significantly for complex microdevices, restricting the application of Maxwell's equations to problems possible to solve. As was illustrated in the previous section, nonlinear electromagnetic and energy conversion phenomena are described by the partial differential equations. The application of Maxwell's equations fulfills the need for data-intensive analysis capabilities with outcome prediction within overall modeling domains as particularly necessary for simulation and analysis of high-performance MEMS. In addition, other modeling and analysis methods are applied. The lumped mathematical models, described by ordinary differential equations, can be used. The process of mathematical modeling and model development is given below.

The first step is to formulate the modeling problem:

- examine and analyze MEMS using a multilevel hierarchy concept, develop multivariable input-output subsystem pairs, e.g., motion microstructures (microscale actuators and sensors), radiating energy microdevices, microscale circuitry, ICs, controller, input/output devices;
- understand and comprehend the MEMS structure and system configuration;
- gather the data and information;
- develop input-output variable pairs, identify the independent and dependent control, disturbance, output, reference (command), state and performance variables, as well as events;

- making accurate assumptions, simplify the problem to make the studied MEMS mathematically tractable (mathematical models, which are the idealization of physical phenomena, are never absolutely accurate, and comprehensive mathematical models simplify the reality to allow the designer to perform a thorough analysis and make accurate predictions of the system performance).

The second step is to derive equations that relate the variables and events:

- define and specify the basic laws (Kirchhoff, Lagrange, Maxwell, Newton, and others) to be used to obtain the equations of motion. Mathematical models of electromagnetic, electronic, and mechanical microscale subsystems can be found and augmented to derive mathematical models of MEMS using defined variables and events;
- derive mathematical models;

The third step is the simulation, analysis, and validation:

- identify the numerical and analytic methods to be used in analysis and simulations;
- analytically and/or numerically solve the mathematical equations (e.g., differential or difference equations, nonlinear equations, etc.);
- using information variables (measured or observed) and events, synthesize the fitting and mismatch functionals;
- verify the results through the comprehensive comparison of the solution (model input-state-output-event mapping sets) with the experimental data (experimental input-state-output-event mapping sets);
- calculate the fitting and mismatch functionals;
- examine the analytical and numerical data against new experimental data and evidence.

If the matching with the desired accuracy is not guaranteed, the mathematical model of MEMS must be refined, and the designer must start the cycle again.

Electromagnetic theory and classical mechanics form the basis for the development of mathematical models of MEMS. It was illustrated that MEMS can be modeled using Maxwell's equations and *torsional-mechanical* equations of motion. However, from modeling, analysis, design, control, and simulation perspectives, the mathematical models as given by ordinary differential equations can be derived and used.

Consider the rotational microstructure (bar magnet, current loop, and microsolenoid) in a uniform magnetic field, see Fig. 14.5. The microstructure rotates if the electromagnetic torque is developed. The electromagnetic field must be studied to find the electromagnetic torque.

The torque tends to align the magnetic moment $\vec{m}$ with $\vec{B}$, and

$$\vec{T} = \vec{m} \times \vec{B}$$

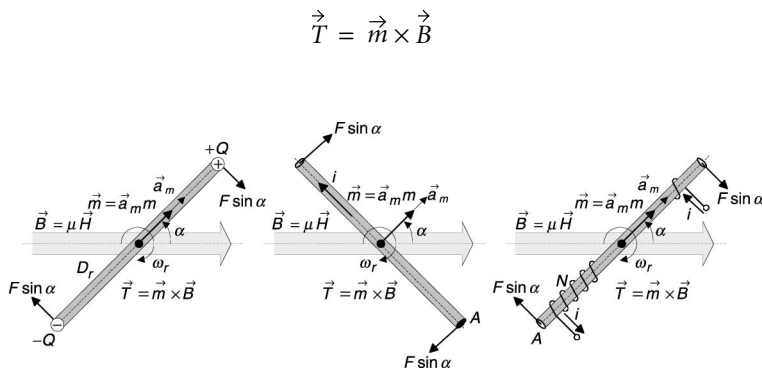



FIGURE 14.5 Clockwise rotation of the motion microstructure.

For a microstructure with outside diameter D_r, the magnet strength is Q. Hence, the magnetic moment is $m = QD_r$, and the force is found as $F = QB$.

The electromagnetic torque is

$$T = 2F\frac{1}{2}D_r \sin\alpha = QD_r B \sin\alpha = mB \sin\alpha$$

Using the unit vector in the magnetic moment direction $\vec{a}_m$, one obtains

$$\vec{T} = \vec{m} \times \vec{B} = \vec{a}_m m \times \vec{B} = QD_r \vec{a}_m \times \vec{B}$$

For a current loop with the area A, the torque is found as

$$\vec{T} = \vec{m} \times \vec{B} = \vec{a}_m m \times \vec{B} = iA\vec{a}_m \times \vec{B}$$

For a solenoid with N turns, one obtains

$$\vec{T} = \vec{m} \times \vec{B} = \vec{a}_m m \times \vec{B} = iAN\vec{a}_m \times \vec{B}$$

As the electromagnetic torque is found, using Newton's second law, one has

$$\frac{d\omega_r}{dt} = \frac{1}{J}\sum \vec{T}_\Sigma = \frac{1}{J}(\vec{T} - \vec{T}_L), \qquad \frac{d\theta_r}{dt} = \omega_r$$

where $\vec{T}_L$ is the load torque.

The *electromotive* (*emf*) and *magnetomotive* (*mmf*) forces can be used in the model development. We have

$$\text{emf} = \oint_l \vec{E} \cdot d\vec{l} = \underbrace{\oint_l (\vec{v} \times \vec{B}) \cdot d\vec{l}}_{\text{motional induction generation}} - \underbrace{\int_s \frac{\partial \vec{B}}{\partial t} d\vec{s}}_{\text{transformer induction}}$$

and

$$\text{mmf} = \int_l \vec{H} \cdot d\vec{l} = \oint_s \vec{J} \cdot d\vec{s} + \oint_s \frac{\partial \vec{D}}{\partial t} d\vec{s}$$

For preliminary design, it is sufficiently accurate to apply Faraday's or Lenz's laws, which give the electromotive force in term of the time-varying magnetic field changes. In particular,

$$\text{emf} = -\frac{d\psi}{dt} = -\frac{\partial \psi}{\partial t} - \frac{\partial \psi}{\partial \theta_r}\frac{d\theta_r}{dt} = -\frac{\partial \psi}{\partial t} - \frac{\partial \psi}{\partial \theta_r}\omega_r$$

where $\frac{\partial \psi}{\partial t}$ is the transformer term.

The total flux linkages are

$$\psi = \frac{1}{4}\pi N_S \Phi_p$$

where N_S is the number of turns and Φ_p is the flux per pole.

For radial topology micromachines, we have

$$\Phi_p = \frac{\mu i N_S}{P^2 g_e} R_{\text{in st}} L$$

where i is the current in the phase microwinding (supplied by the IC), $R_{\text{in st}}$ is the inner stator radius, L is the inductance, P is the number of poles, and g_e is the equivalent gap, which includes the airgap and radial thickness of the permanent magnet.

Denoting the number of turns per phase as N_S, the magnetomotive force is

$$\text{mmf} = \frac{iN_S}{P}\cos P\theta_r$$

The simplified expression for the electromagnetic torque for radial topology brushless micromachines is

$$T = \frac{1}{2}PB_{\text{ag}}i_s N_S L_r D_r$$

where B_{ag} is the air gap flux density, $B_{\text{ag}} = (\mu i N_S/2Pg_e)\cos P\theta_r$, i_s is the total current, L_r is the active length (rotor axial length), and D_r is the outside rotor diameter.

The axial topology brushless micromachines can be designed and fabricated. The electromagnetic torque is given as

$$T = k_{\text{ax}}B_{\text{ag}}i_s N_S D_a^2$$

where k_{ax} is the nonlinear coefficient, which is found in terms of active conductors and thin-film permanent magnet length; and D_a is the equivalent diameter, which is a function of windings and permanent-magnet topography.

Example 14.5.1: Mathematical Model of the Translational Microtransducer

Figure 14.6 illustrates a simple translational microstructure with a stationary member and movable translational microstructure (plunger), which can be fabricated using continuous batch-fabrication process [2]. The winding can be "printed" using the micromachining/CMOS technology.

We apply Newton's second law of motion to study the dynamics. Newton's law states that the acceleration of an object is proportional to the net force. The vector sum of all forces is found as

$$F(t) = m\frac{d^2x}{dt^2} + B_v\frac{dx}{dt} + (k_{s1}x + k_{s2}x^2) + F_e(t)$$

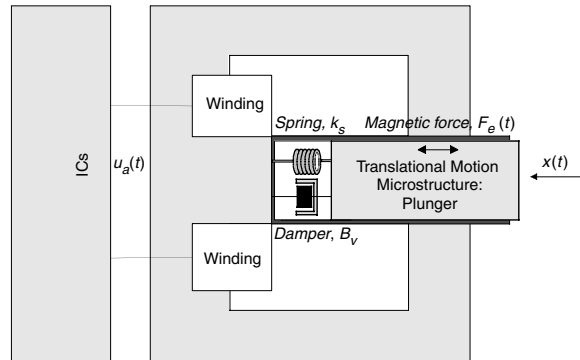



FIGURE 14.6 Microtransducer schematics with translational motion microstructure.

where x is the displacement of a translational microstructure (plunger), m is the mass of a movable plunger, B_v is the viscous friction coefficient, k_{s1} and k_{s2} are the spring constants (the spring can be made from polysilicon), and $F_e(t)$ is the magnetic force which is found using the coenergy W_c, $F_e(i, x) = \frac{\partial W_c(i, x)}{\partial x}$.

The stretch and restoring forces are not directly proportional to the displacement, and these forces are different on either side of the equilibrium position. The restoring/stretching force exerted by the polysilicon spring is expressed as $(k_{s1}x + k_{s2}x^2)$.

Assuming that the magnetic system is linear, the coenergy is expressed as

$$W_c(i, x) = \frac{1}{2}L(x)i^2$$

Then

$$F_e(i, x) = \frac{1}{2}i^2\frac{dL(x)}{dx}$$

The inductance is found as

$$L(x) = \frac{N^2}{\Re_f + \Re_g} = \frac{N^2\mu_f\mu_0A_fA_g}{A_gl_f + 2A_f\mu_f(x + 2d)}$$

where $\Re_f$ and $\Re_g$ are the reluctances of the ferromagnetic material and air gap, A_f and A_g are the associated cross section areas, and l_f and $(x + 2d)$ are the lengths of the magnetic material and the air gap.
Hence

$$\frac{dL}{dx} = -\frac{2N^2\mu_f^2\mu_0A_f^2A_g}{[A_gl_f + 2A_f\mu_f(x + 2d)]^2}$$

Using Kirchhoff's law, the voltage equation for the phase microcircuitry is

$$u_a = ri + \frac{d\psi}{dt}$$

where the flux linkage ψ is expressed as $\psi = L(x)i$.

One obtains

$$u_a = ri + L(x)\frac{di}{dt} + i\frac{dL(x)}{dx}\frac{dx}{dt}$$

and thus

$$\frac{di}{dt} = -\frac{r}{L(x)}i + \frac{2N^2\mu_f^2\mu_0A_f^2A_g}{L(x)[A_gl_f + 2A_f\mu_f(x + 2d)]^2}iv + \frac{1}{L(x)}u_a$$

Augmenting this equation with differential equation

$$F(t) = m\frac{d^2x}{dt^2} + B_v\frac{dx}{dt} + (k_{s1}x + k_{s2}x^2) + F_e(t)$$

three nonlinear differential equations for the studied translation microdevise are found as

$$\frac{di}{dt} = -\frac{r[A_g l_f + 2A_f \mu_f (x+2d)]}{N^2 \mu_f \mu_0 A_f A_g} i + \frac{2\mu_f A_f}{A_g l_f + 2A_f \mu_f (x+2d)} iv + \frac{A_g l_f + 2A_f \mu_f (x+2d)}{N^2 \mu_f \mu_0 A_f A_g} u_a$$

$$\frac{dx}{dt} = v$$

$$\frac{dv}{dt} = \frac{N^2 \mu_f^2 \mu_0 A_f^2 A_g}{m[A_g l_f + 2A_f \mu_f (x+2d)]^2} i^2 - \frac{1}{m}(k_{s1} x + k_{s2} x^2) - \frac{B_v}{m} v$$

Example 14.5.2: Mathematical Model of an Elementary Synchronous Reluctance Micromotor

Consider a single-phase reluctance micromotor, which can be straightforwardly fabricated using conventional CMOS, LIGA, and LIGA-like technologies. Ferromagnetic materials are used to fabricate microscale stator and rotor, and windings can be deposited on the stator, see Fig. 14.7.

The *quadrature* and *direct* magnetic axes are fixed with the microrotor, which rotates with angular velocity ω_r. These magnetic axes rotate with the angular velocity ω. Assume that the initial conditions are zero. Hence, the angular displacements of the rotor θ_r and the angular displacement of the *quadrature* magnetic axis θ are equal, and

$$\theta_r = \theta = \int_{t_0}^{t} \omega_r(\tau) d\tau = \int_{t_0}^{t} \omega(\tau)\, d\tau.$$

The magnetizing reluctance $\Re_m$ is a function of the rotor angular displacement θ_r. Using the number of turns N_S, the magnetizing inductance is

$$L_m(\theta_r) = \frac{N_S^2}{\Re_m(\theta_r)}.$$

This magnetizing inductance varies twice per one revolution of the rotor and has minimum and maximum values, and

$$L_{m\ \min} = \left.\frac{N_S^2}{\Re_{m\ \max}(\theta_r)}\right|_{\theta_r = 0, \pi, 2\pi, \ldots}, \quad L_{m\ \max} = \left.\frac{N_S^2}{\Re_{m\ \min}(\theta_r)}\right|_{\theta_r = \frac{1}{2}\pi, \frac{3}{2}\pi, \frac{5}{2}\pi, \ldots}$$

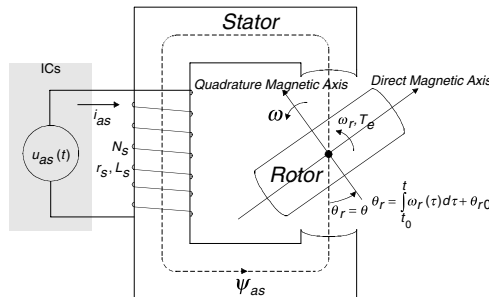



FIGURE 14.7 Microscale single-phase reluctance motor with rotational motion microstructure (microrotor).

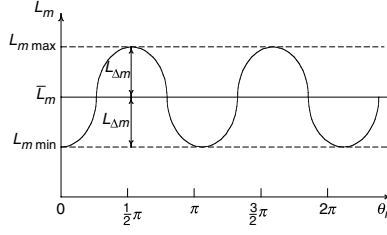



FIGURE 14.8 Magnetizing inductance $L_m(\theta_r)$.

Assume that this variation is a sinusoidal function of the rotor angular displacement. Then,

$$L_m(\theta_r) = \bar{L}_m - L_{\Delta m}\cos 2\theta_r$$

where $\bar{L}_m$ is the average value of the magnetizing inductance and $L_{\Delta m}$ is half of the amplitude of the sinusoidal variation of the magnetizing inductance.

The plot for $L_m(\theta_r)$ is documented in Fig. 14.8.

The electromagnetic torque, developed by single-phase reluctance motors is found using the expression for the coenergy $W_c(i_{as}, \theta_r)$. From $W_c(i_{as}, \theta_r) = \frac{1}{2}(L_{ls} + \bar{L}_m - L_{\Delta m}\cos 2\theta_r)i_{as}^2$, one finds

$$T_e = \frac{\partial W_c(i_{as}, \theta_r)}{\partial \theta_r} = \frac{\partial[\frac{1}{2}i_{as}^2(L_{ls} + \bar{L}_m - L_{\Delta m}\cos 2\theta_r)]}{\partial \theta_r} = L_{\Delta m}i_{as}^2 \sin 2\theta_r$$

The electromagnetic torque is not developed by synchronous reluctance motors if IC feeds the dc current or voltage to the motor winding because $T_e = L_{\Delta m}i_{as}^2 \sin 2\theta_r$. Hence, conventional control algorithms cannot be applied, and new methods, which are based upon electromagnetic features must be researched. The average value of T_e is not equal to zero if the current is a function of θ_r. As an illustration, let us assume that the following current is fed to the motor winding:

$$i_{as} = i_M \operatorname{Re}(\sqrt{\sin 2\theta_r})$$

Then, the electromagnetic torque is

$$T_e = L_{\Delta m}i_{as}^2 \sin 2\theta_r = L_{\Delta m}i_M^2(\operatorname{Re}\sqrt{\sin 2\theta_r})^2 \sin 2\theta_r \neq 0$$

and

$$T_{e\,av} = \frac{1}{\pi}\int_0^{\pi} L_{\Delta m}i_{as}^2 \sin 2\theta_r\, d\theta_r = \frac{1}{4}L_{\Delta m}i_M^2$$

The mathematical model of the microscale single-phase reluctance motor is found by using Kirchhoff's and Newton's second laws

$$u_{as} = r_s i_{as} + \frac{d\psi_{as}}{dt} \quad \text{(circuitry equation)}$$

$$T_e - B_m\omega_r - T_L = J\frac{d^2\theta_r}{dt^2} \quad (\textit{torsional-mechanical} \text{ equation})$$

From $\psi_{as} = (L_{ls} + \bar{L}_m - L_{\Delta m}\cos 2\theta_r)i_{as}$, one obtains a set of three first-order nonlinear differential equations. In particular, we have

$$\frac{di_{as}}{dt} = \frac{r_s}{L_{ls} + \bar{L}_m - L_{\Delta m}\cos 2\theta_r} i_{as} - \frac{2L_{\Delta m}}{L_{ls} + \bar{L}_m - L_{\Delta m}\cos 2\theta_r} i_{as}\omega_r \sin 2\theta_r + \frac{1}{L_{ls} + \bar{L}_m - L_{\Delta m}\cos 2\theta_r} u_{as}$$
$$\frac{d\omega_r}{dt} = \frac{1}{J}(L_{\Delta m} i_{as}^2 \sin 2\theta_r - B_m\omega_r - T_L)$$
$$\frac{d\theta_r}{dt} = \omega_r$$

Example 14.5.3: Mathematical Model of Two-Phase Permanent-Magnet Stepper Micromotors

For two-phase permanent-magnet stepper micromotors, we have

$$u_{as} = r_s i_{as} + \frac{d\psi_{as}}{dt}$$
$$u_{bs} = r_s i_{bs} + \frac{d\psi_{bs}}{dt}$$

where the flux linkages are $\psi_{as} = L_{asas}i_{as} + L_{asbs}i_{bs} + \psi_{asm}$ and $\psi_{bs} = L_{bsas}i_{as} + L_{bsbs}i_{bs} + \psi_{bsm}$.

Here, u_{as} and u_{bs} are the phase voltages in the stator microwindings *as* and *bs*; i_{as} and i_{bs} are the phase currents in the stator microwindings; ψ_{as} and ψ_{bs} are the stator flux linkages; r_s are the resistances of the stator microwindings; L_{asas}, L_{asbs}, L_{bsas}, and L_{bsbs} are the mutual inductances.

The electrical angular velocity and displacement are found using the number of rotor tooth *RT*,

$$\omega_r = RT\omega_{rm}$$
$$\theta_r = RT\theta_{rm}$$

where ω_r and ω_{rm} are the electrical and rotor angular velocities, and θ_r and θ_{rm} are the electrical and rotor angular displacements.

The flux linkages are functions of the number of the rotor tooth *RT*, and the magnitude of the flux linkages produced by the permanent magnets ψ_m. In particular,

$$\psi_{asm} = \psi_m \cos(RT\theta_{rm}) \qquad \text{and} \qquad \psi_{bsm} = \psi_m \sin(RT\theta_{rm})$$

The self-inductance of the stator windings is

$$L_{ss} = L_{asas} = L_{bsbs} = L_{ls} + \bar{L}_m$$

The stator microwindings are displaced by 90 electrical degrees. Hence, the mutual inductances between the stator microwindings are zero, $L_{asbs} = L_{bsas} = 0$.

Then, we have

$$\psi_{as} = L_{ss}i_{as} + \psi_m\cos(RT\theta_{rm}) \qquad \text{and} \qquad \psi_{bs} = L_{ss}i_{bs} + \psi_m\sin(RT\theta_{rm})$$

Taking note of the circuitry equations, one has

$$
\begin{aligned}
u_{as} &= r_s i_{as} + \frac{d[L_{ss} i_{as} + \psi_m \cos(RT\theta_{rm})]}{dt} = r_s i_{as} + L_{ss}\frac{di_{as}}{dt} - RT\psi_m \omega_{rm} \sin(RT\theta_{rm}) \\
u_{bs} &= r_s i_{bs} + \frac{d[L_{ss} i_{bs} + \psi_m \sin(RT\theta_{rm})]}{dt} = r_s i_{bs} + L_{ss}\frac{di_{bs}}{dt} + RT\psi_m \omega_{rm} \cos(RT\theta_{rm})
\end{aligned}
$$

Therefore, we obtain

$$
\begin{aligned}
\frac{di_{as}}{dt} &= -\frac{r_s}{L_{ss}} i_{as} + \frac{RT\psi_m}{L_{ss}} \omega_{rm} \sin(RT\theta_{rm}) + \frac{1}{L_{ss}} u_{as} \\
\frac{di_{bs}}{dt} &= -\frac{r_s}{L_{ss}} i_{bs} - \frac{RT\psi_m}{L_{ss}} \omega_{rm} \cos(RT\theta_{rm}) + \frac{1}{L_{ss}} u_{bs}
\end{aligned}
$$

Using Newton's second law, we have

$$
\begin{aligned}
\frac{d\omega_{rm}}{dt} &= \frac{1}{J}(T_e - B_m \omega_{rm} - T_L) \\
\frac{d\theta_{rm}}{dt} &= \omega_{rm}
\end{aligned}
$$

The expression for the electromagnetic torque developed by permanent-magnet stepper micromotors must be found. Taking note of the relationship for the coenergy

$$
W_c = \frac{1}{2}(L_{ss} i_{as}^2 + L_{ss} i_{bs}^2) + \psi_m i_{as} \cos(RT\theta_{rm}) + \psi_m i_{bs} \sin(RT\theta_{rm}) + W_{PM}
$$

one finds the electromagnetic torque:

$$
T_e = \frac{\partial W_c}{\partial \theta_{rm}} = -RT\psi_m [i_{as} \sin(RT\theta_{rm}) - i_{bs} \cos(RT\theta_{rm})]
$$

Hence, the transient evolution of the phase currents i_{as} and i_{bs}, rotor angular velocity ω_{rm}, and displacement θ_{rm}, is modeled by the following differential equations:

$$
\begin{aligned}
\frac{di_{as}}{dt} &= -\frac{r_s}{L_{ss}} i_{as} + \frac{RT\psi_m}{L_{ss}} \omega_{rm} \sin(RT\theta_{rm}) + \frac{1}{L_{ss}} u_{as} \\
\frac{di_{bs}}{dt} &= -\frac{r_s}{L_{ss}} i_{bs} - \frac{RT\psi_m}{L_{ss}} \omega_{rm} \cos(RT\theta_{rm}) + \frac{1}{L_{ss}} u_{bs} \\
\frac{d\omega_{rm}}{dt} &= -\frac{RT\psi_m}{J} [i_{as} \sin(RT\theta_{rm}) - i_{bs} \cos(RT\theta_{rm})] - \frac{B_m}{J} \omega_{rm} - \frac{1}{J} T_L \\
\frac{d\theta_{rm}}{dt} &= \omega_{rm}
\end{aligned}
$$

These four nonlinear differential equations are rewritten in the state-space form as

$$\begin{bmatrix} \frac{di_{as}}{dt} \\ \frac{di_{bs}}{dt} \\ \frac{d\omega_{rm}}{dt} \\ \frac{d\theta_{rm}}{dt} \end{bmatrix} = \begin{bmatrix} -\frac{r_s}{L_{ss}} & 0 & 0 & 0 \\ 0 & -\frac{r_s}{L_{ss}} & 0 & 0 \\ 0 & 0 & -\frac{B_m}{J} & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} i_{as} \\ i_{bs} \\ \omega_{rm} \\ \theta_{rm} \end{bmatrix} + \begin{bmatrix} \frac{RT\psi_m}{L_{ss}}\omega_{rm}\sin(RT\theta_{rm}) \\ -\frac{RT\psi_m}{L_{ss}}\omega_{rm}\cos(RT\theta_{rm}) \\ -\frac{RT\psi_m}{J}[i_{as}\sin(RT\theta_{rm}) - i_{bs}\cos(RT\theta_{rm})] \\ 0 \end{bmatrix}$$

$$+ \begin{bmatrix} \frac{1}{L_{ss}} & 0 \\ 0 & \frac{1}{L_{ss}} \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_{as} \\ u_{bs} \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ \frac{1}{J} \\ 0 \end{bmatrix} T_L$$

The analysis of the torque equation

$$T_e = -RT\psi_m[i_{as}\sin(RT\theta_{rm}) - i_{bs}\cos(RT\theta_{rm})]$$

guides one to the conclusion that the expressions for a balanced two-phase current sinusoidal set is

$$i_{as} = -\sqrt{2}i_M\sin(RT\theta_{rm}) \quad \text{and} \quad i_{bs} = \sqrt{2}i_M\cos(RT\theta_{rm})$$

If these phase currents are fed, the electromagnetic torque is a function of the current magnitude i_M, and

$$T_e = \sqrt{2}RT\psi_m i_M$$

The phase currents needed to be fed are the functions of the rotor angular displacement. Assuming that the inductances are negligibly small, we have the following phase voltages needed to be supplied:

$$u_{as} = -\sqrt{2}u_M\sin(RT\theta_{rm}) \quad \text{and} \quad u_{bs} = \sqrt{2}u_M\cos(RT\theta_{rm})$$

Example 14.5.4: Mathematical Model of Two-Phase Permanent-Magnet Synchronous Micromotors

Consider two-phase permanent-magnet synchronous micromotors. Using Kirchhoff's voltage law, we have

$$u_{as} = r_s i_{as} + \frac{d\psi_{as}}{dt}$$

$$u_{bs} = r_s i_{bs} + \frac{d\psi_{bs}}{dt}$$

where the flux linkages are expressed as $\psi_{as} = L_{asas}i_{as} + L_{asbs}i_{bs} + \psi_{asm}$ and $\psi_{bs} = L_{bsas}i_{as} + L_{bsbs}i_{bs} + \psi_{bsm}$.

The flux linkages are periodic functions of the angular displacement (rotor position), and let

$$\psi_{asm} = \psi_m\sin\theta_{rm} \quad \text{and} \quad \psi_{bsm} = -\psi_m\cos\theta_{rm}$$

The self-inductances of the stator windings are found to be

$$L_{ss} = L_{asas} = L_{bsbs} = L_{ls} + \bar{L}_m$$

The stator windings are displaced by 90 electrical degrees, and hence, the mutual inductances between the stator windings are $L_{asbs} = L_{bsas} = 0$. Thus, we have

$$\psi_{as} = L_{ss} i_{as} + \psi_m \sin\theta_{rm} \quad \text{and} \quad \psi_{bs} = L_{ss} i_{bs} - \psi_m \cos\theta_{rm}$$

Therefore, one finds

$$\begin{aligned} u_{as} &= r_s i_{as} + \frac{d(L_{ss} i_{as} + \psi_m \sin\theta_{rm})}{dt} = r_s i_{as} + L_{ss}\frac{di_{as}}{dt} + \psi_m \omega_{rm} \cos\theta_{rm} \\ u_{bs} &= r_s i_{bs} + \frac{d(L_{ss} i_{bs} - \psi_m \cos\theta_{rm})}{dt} = r_s i_{bs} + L_{ss}\frac{di_{bs}}{dt} - \psi_m \omega_{rm} \sin\theta_{rm} \end{aligned}$$

Using Newton's second law

$$T_e - B_m \omega_{rm} - T_L = J\frac{d^2\theta_{rm}}{dt^2}$$

we have

$$\begin{aligned} \frac{d\omega_{rm}}{dt} &= \frac{1}{J}(T_e - B_m \omega_{rm} - T_L) \\ \frac{d\theta_{rm}}{dt} &= \omega_{rm} \end{aligned}$$

The expression for the electromagnetic torque developed by permanent-magnet motors can be obtained by using the coenergy

$$W_c = \frac{1}{2}(L_{ss} i_{as}^2 + L_{ss} i_{bs}^2) + \psi_m i_{as} \sin\theta_{rm} - \psi_m i_{bs} \cos\theta_{rm} + W_{PM}$$

Then, one has

$$T_e = \frac{\partial W_c}{\partial \theta_{rm}} = \frac{P\psi_m}{2}(i_{as}\cos\theta_{rm} + i_{bs}\sin\theta_{rm})$$

Augmenting the circuitry transients with the *torsional-mechanical* dynamics, one finds the mathematical model of two-phase permanent-magnet micromotors in the following form:

$$\begin{aligned} \frac{di_{as}}{dt} &= -\frac{r_s}{L_{ss}} i_{as} - \frac{\psi_m}{L_{ss}} \omega_{rm} \cos\theta_{rm} + \frac{1}{L_{ss}} u_{as} \\ \frac{di_{bs}}{dt} &= -\frac{r_s}{L_{ss}} i_{bs} + \frac{\psi_m}{L_{ss}} \omega_{rm} \sin\theta_{rm} + \frac{1}{L_{ss}} u_{bs} \\ \frac{d\omega_{rm}}{dt} &= \frac{P\psi_m}{2J}(i_{as}\cos\theta_{rm} + i_{bs}\sin\theta_{rm}) - \frac{B_m}{J}\omega_{rm} - \frac{1}{J}T_L \\ \frac{d\theta_{rm}}{dt} &= \omega_{rm} \end{aligned}$$

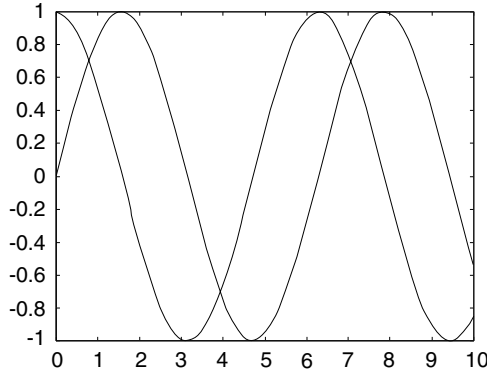



FIGURE 14.9 Air-gap mmf and the phase current waveforms.

For two-phase motors (assuming the sinusoidal winding distributions and the sinusoidal mmf waveforms), the electromagnetic torque is expressed as

$$T_e = \frac{P\psi_m}{2}(i_{as}\cos\theta_{rm} + i_{bs}\sin\theta_{rm})$$

Hence, to guarantee the balanced operation, one feeds

$$i_{as} = \sqrt{2}i_M\cos\theta_{rm} \quad \text{and} \quad i_{bs} = \sqrt{2}i_M\sin\theta_{rm}$$

to maximize the electromagnetic torque. In fact, one obtains

$$T_e = \frac{P\psi_m}{2}(i_{as}\cos\theta_{rm} + i_{bs}\sin\theta_{rm}) = \frac{P\psi_m}{2}\sqrt{2}i_M(\cos^2\theta_{rm} + \sin^2\theta_{rm}) = \frac{P\psi_m}{\sqrt{2}}i_M$$

The air-gap mmf and the phase current waveforms are plotted in Fig. 14.9.

14.6 Control of MEMS

Mathematical models of MEMS can be developed with different degrees of complexity. It must be emphasized that in addition to the models of microscale motion devices, the fast dynamics of ICs should be examined. Due to the complexity of complete mathematical models of ICs, impracticality of the developed equations, and very fast dynamics, the IC dynamics can be modeled using reduced-order differential equation or as unmodeled dynamics. For MEMS, modeled using linear and nonlinear differential equations

$$\dot{x}(t) = Ax + Bu, \quad u_{\min} \le u \le u_{\max}, \quad y = Hx$$

$$\dot{x}(t) = F_z(t, x, r, z) + B_p(t, x, p)u, \quad u_{\min} \le u \le u_{\max}, \quad y = H(x)$$

different control algorithms can be designed.

Here, the state, control, output, and reference (command) vectors are denoted as x, u, y, and r; parameter uncertainties (e.g., time-varying coefficients, unmodeled dynamics, unpredicted changes, etc.) are modeled using z and p vectors.

The matrices of coefficients are A, B, and H. The smooth mapping fields of the nonlinear model are denoted as $F_z(\cdot)$, $B_p(\cdot)$, and $H(\cdot)$.

It should be emphasized that the control is bounded. For example, using the IC duty ratio d_D as the control signal, we have $0 \le d_D \le 1$ or $-1 \le d_D \le +1$. Four-quadrant ICs are used due to superior performance, and $-1 \le d_D \le +1$. Hence, we have $-1 \le u \le +1$. However, in general, $u_{\min} \le u \le u_{\max}$.

Proportional-Integral-Derivative Control

Many MEMS can be controlled by the proportional-integral-derivative (PID) controllers, which, taking note of control bounds, are given as [9]

$$u(t) = \mathbf{sat}_{u_{\min}}^{u_{\max}}\left(e, \int e\,dt, \frac{de}{dt}\right)$$

$$= \mathbf{sat}_{u_{\min}}^{u_{\max}}\left(\underbrace{\sum_{j=0}^{\varsigma} k_{pj} e^{\frac{2j+1}{2\beta+1}}}_{\text{proportional}} + \underbrace{\sum_{j=0}^{\sigma} k_{ij} \int e^{\frac{2j+1}{2\mu+1}} dt}_{\text{integral}} + \underbrace{\sum_{j=0}^{\alpha} k_{dj} \dot{e}^{\frac{2j+1}{2\gamma+1}}}_{\text{derivative}}\right), \quad u_{\min} \le u \le u_{\max}$$

where k_{pj}, k_{ij}, and k_{dj} are the matrices of the proportional, integral, and derivative feedback gains; ς, β, σ, μ, α, and γ are the nonnegative integers.

In the nonlinear PID controllers, the tracking error is used. In particular,

$$e(t) = \underbrace{r(t)}_{\text{reference/command}} - \underbrace{y(t)}_{\text{output}}$$

Linear bounded controllers can be straightforwardly designed. For example, letting $\varsigma = \beta = \sigma = \mu = 0$, we have the following linear PI control law:

$$u(t) = \mathbf{sat}_{u_{\min}}^{u_{\max}}\left(k_{p0} e(t) + k_{i0} \int et\,dt\right)$$

The PID controllers with the state feedback extension can be synthesized as

$$u(t) = \mathbf{sat}_{u_{\min}}^{u_{\max}}(e, x)$$

$$= \mathbf{sat}_{u_{\min}}^{u_{\max}}\left(\underbrace{\sum_{j=0}^{\varsigma} k_{pj} e^{\frac{2j+1}{2\beta+1}}}_{\text{proportional}} + \underbrace{\sum_{j=0}^{\sigma} k_{ij} \int e^{\frac{2j+1}{2\mu+1}} dt}_{\text{integral}} + \underbrace{\sum_{j=0}^{\alpha} k_{dj} \dot{e}^{\frac{2j+1}{2\gamma+1}}}_{\text{derivative}} + G(t) B \frac{\partial V(e, x)}{\partial \begin{bmatrix} e \\ x \end{bmatrix}}\right), \quad u_{\min} \le u \le u_{\max}$$

where $V(e, x)$ is the function that satisfies the general requirements imposed on the Lyapunov pair [9], e.g., the sufficient conditions for stability are used.

It is evident that nonlinear feedback mappings result, and the nonquadratic function $V(e, x)$ can be synthesized and used to obtain the control algorithm and feedback gains.

Tracking Control

Tracking control is designed for the augmented systems, which are modeled using the state variables and the reference dynamics. In particular, from

$$\dot{x}(t) = Ax + Bu, \quad \dot{x}^{\text{ref}}(t) = r(t) - y(t) = r(t) - Hx(t)$$

one finds

$$\dot{x}_\Sigma(t) = A_\Sigma x_\Sigma + B_\Sigma u + N_\Sigma r, \quad y = Hx, \quad x_\Sigma = \begin{bmatrix} x \\ x^{\text{ref}} \end{bmatrix}, \quad A_\Sigma = \begin{bmatrix} A & 0 \\ -H & 0 \end{bmatrix}, \quad B_\Sigma = \begin{bmatrix} B \\ 0 \end{bmatrix}, \quad N_\Sigma = \begin{bmatrix} 0 \\ I \end{bmatrix}$$

Minimizing the quadratic performance functional

$$J = \frac{1}{2}\int_{t_0}^{t_f} (x_\Sigma^T Q x_\Sigma + u^T G u)\,dt$$

one finds the control law using the first-order necessary condition for optimality. In particular, we have

$$u = -G^{-1} B_\Sigma^T \frac{\partial V}{\partial x_\Sigma} = -G^{-1}\begin{bmatrix} B \\ 0 \end{bmatrix}^T \frac{\partial V}{\partial x_\Sigma}$$

Here, Q is the positive semi-definite constant-coefficient matrix, and G is the positive weighting constant-coefficient matrix.

The solution of the Hamilton–Jacobi equation

$$-\frac{\partial V}{\partial t} = \frac{1}{2} x_\Sigma^T Q x_\Sigma + \left(\frac{\partial V}{\partial x_\Sigma}\right)^T A x_\Sigma - \frac{1}{2}\left(\frac{\partial V}{\partial x_\Sigma}\right)^T B_\Sigma G^{-1} B_\Sigma^T \frac{\partial V}{\partial x_\Sigma}$$

is satisfied by the quadratic return function $V = \frac{1}{2} x_\Sigma^T K x_\Sigma$. Here, K is the symmetric matrix, which must be found by solving the nonlinear differential equation

$$-\dot{K} = Q + A_\Sigma^T K + K^T A_\Sigma - K^T B_\Sigma G^{-1} B_\Sigma^T K, \quad K(t_f) = K_f$$

The controller is given as

$$u = -G^{-1} B_\Sigma^T K x_\Sigma = -G^{-1}\begin{bmatrix} B \\ 0 \end{bmatrix}^T K x_\Sigma$$

From $\dot{x}_{\text{ref}}(t) = e(t)$, one has

$$x_{\text{ref}}(t) = \int e(t)\,dt$$

Therefore, we obtain the integral control law

$$u(t) = -G^{-1}\begin{bmatrix} B \\ 0 \end{bmatrix}^T K \begin{bmatrix} x(t) \\ \int e(t)\,dt \end{bmatrix}$$

In this control algorithm, the error vector is used in addition to the state feedback.

As was illustrated, the bounds are imposed on the control, and $u_{\min} \le u \le u_{\max}$. Therefore, the bounded controllers must be designed. Using the nonquadratic performance functional [9]

$$J = \int_{t_0}^{t_f} \left(x_\Sigma^T Q x_\Sigma + G \int \tan^{-1} u\, du \right) dt$$

with positive semi-definite constant-coefficient matrix Q and positive-definite matrix G, one finds

$$u(t) = -\tanh\left(G^{-1}\begin{bmatrix} B \\ 0 \end{bmatrix}^T K \begin{bmatrix} x(t) \\ \int e(t)dt \end{bmatrix}\right) \approx -\text{sat}_{-1}^{+1}\left(G^{-1}\begin{bmatrix} B \\ 0 \end{bmatrix}^T K \begin{bmatrix} x(t) \\ \int e(t)dt \end{bmatrix}\right), \quad -1 \le u \le 1$$

This controller is obtained assuming that the solution of the functional partial differential equation can be approximated by the quadratic return function

$$V = \frac{1}{2} x_\Sigma^T K x_\Sigma$$

where K is the symmetric matrix.

Time-Optimal Control

A time-optimal controller can be designed using the functional

$$J = \frac{1}{2}\int_{t_0}^{t_f} (x_\Sigma^T Q x_\Sigma)\, dt$$

Taking note of the Hamilton–Jacobi equation

$$-\frac{\partial V}{\partial t} = \min_{-1 \le u \le 1}\left[\frac{1}{2} x_\Sigma^T Q x_\Sigma + \left(\frac{\partial V}{\partial x_\Sigma}\right)^T (A x_\Sigma + B_\Sigma u)\right]$$

the relay-type controller is found to be

$$u = -\text{sgn}\left(B_\Sigma^T \frac{\partial V}{\partial x_\Sigma}\right), \quad -1 \le u \le 1$$

This "optimal" control algorithm cannot be implemented in practice due to the chattering phenomenon. Therefore, relay-type control laws with dead zone

$$u = -\text{sgn}\left(B_\Sigma^T \frac{\partial V}{\partial x_\Sigma}\right)\Bigg|_{\text{dead zone}}, \quad -1 \le u \le 1$$

are commonly used.

Sliding Mode Control

Soft-switching sliding mode control laws are synthesized in [9]. Sliding mode soft-switching algorithms provide superior performance, and the chattering effect is eliminated.

To design controllers, we model the states and errors dynamics as

$$\dot{x}(t) = Ax + Bu, \quad -1 \le u \le 1$$
$$\dot{e}(t) = N\dot{r}(t) - HAx - HBu$$

The smooth sliding manifold is

$$M = \{(t, x, e) \in R_{\ge 0} \times X \times E \mid \upsilon(t, x, e) = 0\}$$
$$= \bigcap_{j=1}^{m} \{(t, x, e) \in R_{\ge 0} \times X \times E \mid \upsilon_j(t, x, e) = 0\}$$

The time-varying nonlinear switching surface is $\upsilon(t, x, e) = K_{\upsilon xe}(t, x, e) = 0$. The soft-switching control law is given as

$$u(t, x, e) = -G\phi(\upsilon), \quad -1 \le u \le 1, \quad G > 0$$

where $\phi(\cdot)$ is the continuous real-analytic function of class $C^{\epsilon}(\epsilon \ge 1)$, for example, `tanh` and `erf`.

Constrained Control of Nonlinear MEMS: Hamilton–Jacobi Method

Constrained optimization of MEMS is a topic of great practical interest. Using the Hamilton–Jacobi theory, the bounded controllers can be synthesized for continuous-time systems modeled as

$$\dot{x}^{\mathrm{MEMS}}(t) = F_s(x^{\mathrm{MEMS}}) + B_s(x^{\mathrm{MEMS}})u^{2w+1}, \qquad y = Hx^{\mathrm{MEMS}}$$
$$u_{\min} \le u \le u_{\max}, \quad x^{\mathrm{MEMS}}(t_0) = x_0^{\mathrm{MEMS}}$$

Here, $x^{\mathrm{MEMS}} \in X_s$ is the state vector; $u \in U$ is the vector of control inputs: $y \in Y$ is the measured output; $F_s(\cdot)$, $B_s(\cdot)$ and $H(\cdot)$ are the smooth mappings; $F_s(0) = 0$, $B_s(0) = 0$, and $H(0) = 0$; and w is the nonnegative integer.

To design the tracking controller, we augment the MEMS dynamics

$$\dot{x}^{\mathrm{MEMS}}(t) = F_s(x^{\mathrm{MEMS}}) + B_s(x^{\mathrm{MEMS}})u^{2w+1} \qquad y = H(x^{\mathrm{MEMS}})$$
$$u_{\min} \le u \le u_{\max}, \quad x^{\mathrm{MEMS}}(t_0) = x_0^{\mathrm{MEMS}}$$

with the *exogenous* dynamics $\dot{x}^{\mathrm{ref}}(t) = Nr - y = Nr - H(x^{\mathrm{MEMS}})$.

Using the augmented state vector

$$x = \begin{bmatrix} x^{\mathrm{MEMS}} \\ x^{\mathrm{ref}} \end{bmatrix} \in X$$

one obtains

$$\dot{x}(t) = F(x, r) + B(x)u^{2w+1}, \quad u_{\min} \le u \le u_{\max}, \quad x(t_0) = x_0, \quad x = \begin{bmatrix} x^{\mathrm{MEMS}} \\ x^{\mathrm{ref}} \end{bmatrix}$$

$$F(x, r) = \begin{bmatrix} F_s(x^{\mathrm{MEMS}}) \\ -H(x^{\mathrm{MEMS}}) \end{bmatrix} + \begin{bmatrix} 0 \\ N \end{bmatrix} r, \qquad B(x) = \begin{bmatrix} B_s(x^{\mathrm{MEMS}}) \\ 0 \end{bmatrix}$$

The set of admissible control U consists of the Lebesgue measurable function $u(\cdot)$, and a bounded controller should be designed within the constrained control set

$$U = \{u \in \mathbb{R}^m | u_{i\min} \le u_i \le u_{i\max}, \quad i = 1, \ldots, m\}.$$

We map the control bounds imposed by a bounded, integrable, one-to-one, globally Lipschitz, vector-valued continuous function $\Phi \in C^{\epsilon}(\epsilon \ge 1)$. Our goal is to analytically design the bounded admissible state-feedback controller in the closed form as $u = \Phi(x)$. The most common Φ are the algebraic and transcendental (exponential, hyperbolic, logarithmic, trigonometric) continuously differentiable, integrable, one-to-one functions. For example, the odd one-to-one integrable function tanh with domain $(-\infty, +\infty)$ maps the control bounds. This function has the corresponding inverse function $\tanh^{-1}$ with range $(-\infty, +\infty)$.

The performance cost to be minimized is given as

$$J = \int_{t_0}^{\infty} [W_x(x) + W_u(u)]dt = \int_{t_0}^{\infty} \left[W_x(x) + (2w+1)\int (\Phi^{-1}(u))^T G^{-1} \mathrm{diag}(u^{2w}) du\right] dt$$

where $G^{-1} \in \mathbb{R}^{m \times m}$ is the positive-definite diagonal matrix.

Performance integrands $W_x(\cdot)$ and $W_u(\cdot)$ are real-valued, positive-definite, and continuously differentiable integrand functions. Using the properties of Φ one concludes that inverse function Φ^{-1} is integrable. Hence, integral

$$\int (\Phi^{-1}(u))^T G^{-1} \operatorname{diag}(u^{2w}) du$$

exists.

Example

Consider a nonlinear dynamic system

$$\frac{dx}{dt} = ax + bu^3, \quad u_{\min} \leq u \leq u_{\max}$$

Taking note of

$$W_u(u) = (2w+1)\int (\Phi^{-1}(u))^T G^{-1} \operatorname{diag}(u^{2w}) du$$

one has the positive-definite integrand

$$W_u(u) = 3\int \tanh^{-1} u G^{-1} u^2 du = \frac{1}{3}u^3 \tanh^{-1} u + \frac{1}{6}u^2 + \frac{1}{6}\ln(1-u^2), \quad G^{-1} = \frac{1}{3}$$

In general, if the hyperbolic tangent is used to map the saturation effect, for the single-input case, one has

$$W_u(u) = (2w+1)\int u^{2w} \tanh^{-1}\frac{u}{k} du = u^{2w+1}\tanh^{-1}\frac{u}{k} - k\int \frac{u^{2w+1}}{k^2 - u^2} du$$

Necessary conditions that the control function $u(\cdot)$ guarantees a minimum to the Hamiltonian

$$H = W_x(x) + (2w+1)\int (\Phi^{-1}(u))^T G^{-1} \operatorname{diag}(u^{2w}) du + \frac{\partial V(x)^T}{\partial x}[F(x, r) + B(x)u^{2w+1}]$$

are: first-order necessary condition $n1$,

$$\frac{\partial H}{\partial u} = 0$$

and second-order necessary condition $n2$,

$$\frac{\partial^2 H}{\partial u \times \partial u^T} > 0$$

The positive-definite return function $V(\cdot)$, $V \in C^{\kappa}$, $\kappa \geq 1$, is

$$V(x_0) = \inf_{u \in U} J(x_0, u) = \inf J(x_0, \Phi(\cdot)) \geq 0$$

The Hamilton–Jacobi–Bellman equation is given as

$$-\frac{\partial V}{\partial t} = \min_{u \in U}\left\{W_x(x) + (2w+1)\int (\Phi^{-1}(u))^T G^{-1} \operatorname{diag}(u^{2w}) du + \frac{\partial V(x)^T}{\partial x}[F(x, r) + B(x)u^{2w+1}]\right\}$$

The controller should be derived by finding the control value that attains the minimum to nonquadratic functional. The first-order necessary condition ($n1$) leads us to an admissible bounded control law. In particular,

$$u = -\Phi\left(GB(x)^T \frac{\partial V(x)}{\partial x}\right), \quad u \in U$$

The second-order necessary condition for optimality ($n2$) is met because the matrix G^{-1} is positive-definite. Hence, a unique, bounded, real-analytic, and continuous control candidate is designed.

If there exists a proper function $V(x)$ which satisfies the Hamilton–Jacobi equation, the resulting closed-loop system is robustly stable in the specified state X and control U sets, and robust tracking is ensured in the convex and compact set $XY(X_0, U, R, E_0)$. That is, there exists an invariant domain of stability

$$S = \{x \in \mathbb{R}^c,\ e \in \mathbb{R}^b\colon \|x(t)\| \le \varrho_x(\|x_0\|, t) + \varrho_u(\|u\|),\ \|e(t)\| \le \varrho_e(\|e_0\|, t) + \varrho_r(\|r\|) + \varrho_y(\|y\|),$$
$$\forall x \in X(X_0, U),\ \forall t \in [t_0, \infty),\ \forall e \in E(E_0, R, Y)\} \subset \mathbb{R}^c \times \mathbb{R}^b,$$

and control $u(\cdot)$, $u \in U$ steers the tracking error to the set

$$S_E(\delta) = \{e \in \mathbb{R}^b\colon e_0 \in E_0,\ x \in X(X_0, U),\ r \in R,\ y \in Y,\ t \in [t_0, \infty)$$
$$|\ \|e(t)\| \le \varrho_e(\|e_0\|, t) + \delta,\ \delta \ge 0,\ \forall e \in E(E_0, R, Y),\ \forall t \in [t_0, \infty)\} \subset \mathbb{R}^b$$

Here ϱ_x and ϱ_e are the KL-functions; and ϱ_u, ϱ_r, and ϱ_y are the K-functions.

The solution of the functional equation should be found using nonquadratic return functions. To obtain $V(\cdot)$, the performance cost must be evaluated at the allowed values of the states and control. Linear and nonlinear functionals admit the final values, and the minimum value of the nonquadratic cost is given by power-series forms [9]. That is,

$$J_{\min} = \sum_{i=0}^{\eta} v(x_0)^{\frac{2(i+\gamma+1)}{2\gamma+1}}, \quad \eta = 0, 1, 2, \ldots, \quad \gamma = 0, 1, 2, \ldots$$

The solution of the partial differential equation is satisfied by a continuously differentiable positive-definite return function

$$V(x) = \sum_{i=0}^{\eta} \frac{2\gamma+1}{2(i+\gamma+1)} \left(x^{\frac{i+\gamma+1}{2\gamma+1}}\right)^T K_i x^{\frac{i+\gamma+1}{2\gamma+1}}$$

where matrices K_i are found by solving the Hamilton–Jacobi equation.

The quadratic return function in $V(x) = \frac{1}{2} x^T K_0 x$ is found by letting $\eta = \gamma = 0$. This quadratic candidate may be employed only if the designer enables to neglect the high-order terms in Taylor's series expansion. Using $\eta = 1$ and $\gamma = 0$, one obtains

$$V(x) = \frac{1}{2} x^T K_0 x + \frac{1}{4} (x^2)^T K_1 x^2$$

while for $\eta = 4$ and $\gamma = 1$, we have the following function:

$$V(x) = \frac{3}{4}(x^{2/3})^T K_0 x^{2/3} + \frac{1}{2} x^T K_1 x + \frac{3}{8}(x^{4/3})^T K_2 x^{4/3} + \frac{3}{10}(x^{5/3})^T K_3 x^{5/3} + \frac{1}{4}(x^2)^T K_4 x^2$$

The nonlinear bounded controller is given as

$$u = -\Phi\left(GB(x)^T\sum_{i=0}^{\eta}\operatorname{diag}\left[x(t)^{\frac{i-\gamma}{2\gamma+1}}\right]K_i(t)x(t)^{\frac{i+\gamma+1}{2\gamma+1}}\right),$$

$$\operatorname{diag}\left[x(t)^{\frac{i-\gamma}{2\gamma+1}}\right] = \begin{bmatrix} x_1^{\frac{i-\gamma}{2\gamma+1}} & 0 & \cdots & 0 & 0 \\ 0 & x_2^{\frac{i-\gamma}{2\gamma+1}} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & x_{c-1}^{\frac{i-\gamma}{2\gamma+1}} & 0 \\ 0 & 0 & \vdots & 0 & x_c^{\frac{i-\gamma}{2\gamma+1}} \end{bmatrix}$$

If matrices K_i are diagonal, we have the following control algorithm:

$$u = -\Phi\left(GB(x)^T\sum_{i=0}^{\eta}K_i x^{\frac{2i+1}{2\gamma+1}}\right)$$

Constrained Control of Nonlinear Uncertain MEMS: Lyapunov Method

Over the horizon $[t_0, \infty)$ we consider the dynamics of MEMS modeled as

$$\dot{x}(t) = F_z(t, x, r, z) + B_p(t, x, p)u, \qquad y = H(x), \qquad u_{\min} \le u \le u_{\max}, \quad x(t_0) = x_0$$

where $t \in \mathbb{R}_{\ge 0}$ is the time; $x \in X$ is the state-space vector; $u \in U$ is the vector of bounded control inputs; $r \in R$ and $y \in Y$ are the measured reference and output vectors; $z \in Z$ and $p \in P$ are the parameter uncertainties, functions $z(\cdot)$ and $p(\cdot)$ are Lebesgue measurable and known within bounds; Z and P are the known nonempty compact sets; and $F_z(\cdot)$, $B_p(\cdot)$, and $H(\cdot)$ are the smooth mapping fields.

Let us formulate and solve the motion control problem by synthesizing robust controllers that guarantee stability and robust tracking. Our goal is to design control laws that robustly stabilize nonlinear systems with uncertain parameters and drive the tracking error $e(t) = r(t) - y(t)$, $e \in E$ robustly to the compact set. For MEMS modeled by nonlinear differential equations with parameter variations, the robust tracking of the measured output vector $y \in Y$ must be accomplished with respect to the measured uniformly bounded reference input vector $r \in R$.

The *nominal* and uncertain dynamics are mapped by $F(\cdot)$, $B(\cdot)$, and $\Xi(\cdot)$. Hence, the system evolution is described as

$$\dot{x}(t) = F(t, x, r) + B(t, x)u + \Xi(t, x, u, z, p), \qquad y = H(x), \qquad u_{\min} \le u \le u_{\max}, \quad x(t_0) = x_0$$

There exists a norm of $\Xi(t, x, u, z, p)$, and $\|\Xi(t, x, u, z, p)\| \le \rho(t, x)$, where $\rho(\cdot)$ is the continuous Lebesgue measurable function. Our goal is to solve the motion control problem, and tracking controllers must be synthesized using the tracking error vector and the state variables. Furthermore, to guarantee robustness and to expand stability margins, to improve dynamic performance, and to meet other requirements, nonqudratic Lyapunov functions $V(t, e, x)$ will be used in stability analysis and design of robust tracking control laws.

Suppose that a set of admissible control U consists of the Lebesgue measurable function $u(\cdot)$. It was demonstrated that the Hamilton–Jacobi theory can be used to find control laws, and the minimization of nonquadratic performance functionals leads one to the bounded controllers.

Letting $u = \Phi(t, e, x)$, one obtains a set of admissible controllers. Applying the error and state feedback we define a family of tracking controllers as

$$u = \Omega(x)\Phi(t,e,x) = -\Omega(x)\Phi\left(G_E(t)B_E(t,x)^T\frac{1}{s}\frac{\partial V(t,e,x)}{\partial e} + G_X(t)B(t,x)^T\frac{\partial V(t,e,x)}{\partial x}\right), \quad s = \frac{d}{dt}$$

where $\Omega(\cdot)$ is the nonlinear function; $G_E(\cdot)$ and $G_X(\cdot)$ are the diagonal matrix-functions defined on $[t_0, \infty)$; $B_E(\cdot)$ is the matrix-function; and $V(\cdot)$ is the continuous, differentiable, and real-analytic function.

Let us design the Lyapunov function. This problem is a critical one and involves well-known difficulties. The quadratic Lyapunov candidates can be used. However, for uncertain nonlinear systems, nonquadratic functions $V(t, e, x)$ allow one to realize the full potential of the Lyapunov-based theory and lead us to the nonlinear feedback maps which are needed to achieve conflicting design objectives. We introduce the following family of Lyapunov candidates:

$$V(t,e,x) = \sum_{i=0}^{\varsigma}\frac{2\beta+1}{2(i+\beta+1)}\left(e^{\frac{i+\beta+1}{2\beta+1}}\right)^T K_{Ei}(t)e^{\frac{i+\beta+1}{2\beta+1}} + \sum_{i=0}^{\eta}\frac{2\gamma+1}{2(i+\gamma+1)}\left(x^{\frac{i+\gamma+1}{2\gamma+1}}\right)^T K_{Xi}(t)x^{\frac{i+\gamma+1}{2\gamma+1}}$$

where $K_{Ei}(\cdot)$ and $K_{Xi}(\cdot)$ are the symmetric matrices; ζ, β, η, and γ are the nonnegative integers; $\zeta = 0, 1, 2, \ldots$; $\beta = 0, 1, 2, \ldots$; $\eta = 0, 1, 2, \ldots$; and $\gamma = 0, 1, 2, \ldots$

The well-known quadratic form of $V(t, e, x)$ is found by letting $\zeta = \beta = \eta = \gamma = 0$, and we have

$$V(t,e,x) = \frac{1}{2}e^T K_{E0}(t)e + \frac{1}{2}x^T K_{X0}(t)x$$

By using $\zeta = 1$, $\beta = 0$, $\eta = 1$, and $\gamma = 0$, one obtains a nonquadratic candidate:

$$V(t,e,x) = \frac{1}{2}e^T K_{E0}(t)e + \frac{1}{4}e^{2^T} K_{E1}(t)e^2 + \frac{1}{2}x^T K_{X0}(t)x + \frac{1}{4}x^{2^T} K_{X1}(t)x^2$$

One obtains the following tracking control law:

$$u = -\Omega(x)\Phi\left(G_E(t)B_E(t,x)^T\sum_{i=0}^{\varsigma}\text{diag}\left[e(t)^{\frac{i-\beta}{2\beta+1}}\right]K_{Ei}(t)\frac{1}{s}e(t)^{\frac{i+\beta+1}{2\beta+1}}\right.$$
$$\left. + G_X(t)B(t,x)^T\sum_{i=0}^{\eta}\text{diag}\left[x(t)^{\frac{i-\gamma}{2\gamma+1}}\right]K_{Xi}(t)x(t)^{\frac{i+\gamma+1}{2\gamma+1}}\right)$$

$$\text{diag}\left[e(t)^{\frac{i-\beta}{2\beta+1}}\right] = \begin{bmatrix} e_1^{\frac{i-\beta}{2\beta+1}} & 0 & \cdots & 0 & 0 \\ 0 & e_2^{\frac{i-\beta}{2\beta+1}} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & e_{b-1}^{\frac{i-\beta}{2\beta+1}} & 0 \\ 0 & 0 & \vdots & 0 & e_b^{\frac{i-\beta}{2\beta+1}} \end{bmatrix}$$

and

$$\text{diag}\left[x(t)^{\frac{i-\gamma}{2\gamma+1}}\right] = \begin{bmatrix} x_1^{\frac{i-\gamma}{2\gamma+1}} & 0 & \cdots & 0 & 0 \\ 0 & x_2^{\frac{i-\gamma}{2\gamma+1}} & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & x_{n-1}^{\frac{i-\gamma}{2\gamma+1}} & 0 \\ 0 & 0 & \vdots & 0 & x_n^{\frac{i-\gamma}{2\gamma+1}} \end{bmatrix}$$

If matrices K_{Ei} and K_{Xi} are diagonal, we have

$$u = -\Omega(x)\Phi\left(G_E(t)B_E(t,x)^T\sum_{i=0}^{\varsigma}K_{Ei}(t)\frac{1}{s}e(t)^{\frac{2i+1}{2\beta+1}} + G_X(t)B(t,x)^T\sum_{i=0}^{\eta}K_{Xi}(t)x(t)^{\frac{2i+1}{2\gamma+1}}\right)$$

A closed-loop uncertain system is robustly stable in $X(X_0, U, Z, P)$ and robust tracking is guaranteed in the convex and compact set $E(E_0, Y, R)$ if for reference inputs $r \in R$ and uncertainties in Z and P there exists a $C^\kappa(\kappa \geq 1)$ function $V(\cdot)$, as well as K_∞-functions $\rho_{X1}(\cdot)$, $\rho_{X2}(\cdot)$, $\rho_{E1}(\cdot)$, $\rho_{E2}(\cdot)$ and K-functions $\rho_{X3}(\cdot)$, $\rho_{E3}(\cdot)$, such that the following sufficient conditions:

$$\rho_{X1}(\|x\|) + \rho_{E1}(\|e\|) \leq V(t,e,x) \leq \rho_{X2}(\|x\|) + \rho_{E2}(\|e\|)$$
$$\frac{dV(t,e,x)}{dt} \leq -\rho_{X3}(\|x\|) - \rho_{E3}(\|e\|)$$

are guaranteed in an invariant domain of stability S, and $XE(X_0, E_0, U, R, Z, P) \subseteq S$.

The sufficient conditions under which the robust control problem is solvable were given. Computing the derivative of the $V(t,e,x)$, the unknown coefficients of $V(t,e,x)$ can be found. That is, matrices $K_{Ei}(\cdot)$ and $K_{Xi}(\cdot)$ are obtained. This problem is solved using the nonlinear inequality concept [9].

Example 14.6.1: Control of Two-Phase Permanent-Magnet Stepper Micromotors

High-performance MEMS with permanent-magnet stepper micromotors have been designed and manufactured. Controllers are needed to be designed to control permanent-magnet stepper micromotors, and the angular velocity and position are regulated by changing the magnitude of the voltages applied or currents fed to the stator windings (see Example 14.5.3). The rotor displacement is measured or observed in order to properly apply the voltages to the phase windings. To solve the motion control problem, the controller must be designed. It is illustrated that novel control algorithms are needed to be deployed to maximize the torque developed. In fact, conventional controllers

$$u = -G^{-1}B^T\frac{\partial V}{\partial x} \quad \text{and} \quad u = -\Phi\left(G^{-1}B^T\frac{\partial V}{\partial x}\right)$$

cannot be used.

Using the coenergy concept, one finds the expression for the electromagnetic torque as given by

$$T_e = -RT\psi_m[i_{as}\sin(RT\theta_{rm}) - i_{bs}\cos(RT\theta_{rm})]$$

and thus, one must fed the phase currents as sinusoidal and cosinusoidal functions of the rotor displacement.

The mathematical model of permanent-magnet stepper micromotor was found in Example 14.5.3 as

$$\frac{di_{as}}{dt} = -\frac{r_s}{L_{ss}}i_{as} + \frac{RT\psi_m}{L_{ss}}\omega_{rm}\sin(RT\theta_{rm}) + \frac{1}{L_{ss}}u_{as}$$

$$\frac{di_{bs}}{dt} = -\frac{r_s}{L_{ss}}i_{bs} - \frac{RT\psi_m}{L_{ss}}\omega_{rm}\cos(RT\theta_{rm}) + \frac{1}{L_{ss}}u_{bs}$$

$$\frac{d\omega_{rm}}{dt} = -\frac{RT\psi_m}{J}[i_{as}\sin(RT\theta_{rm}) - i_{bs}\cos(RT\theta_{rm})] - \frac{B_m}{J}\omega_{rm} - \frac{1}{J}T_L$$

$$\frac{d\theta_{rm}}{dt} = \omega_{rm}$$

The rotor resistance is a function of temperature because the resistivity is $\rho_T = \rho_0[(1 + \alpha_\rho(T^\circ - 20^\circ)]$. Hence, $r_s(\cdot) \in [r_{s\min}\ r_{s\max}]$. The susceptibility of the permanent magnets (thin films) decreases with increasing temperature. Other servo-system parameters also vary; in particular, $L_{ss}(\cdot) \in [L_{ss\min}\ L_{ss\max}]$ and $B_m(\cdot) \in [B_{m\min}\ B_{m\max}]$.

The following equation of motion in vector form results:

$$\dot{x}(t) = F_z(t, x, r, d, z) + B_p(p)u, \quad u_{\min} \le u \le u_{\max}$$

$$x(t_0) = x_0, \quad x = \begin{bmatrix} i_{as} \\ i_{bs} \\ \omega_{rm} \\ \theta_{rm} \end{bmatrix}, \quad u = \begin{bmatrix} u_{as} \\ u_{bs} \end{bmatrix}, \quad y = \theta_{rm}$$

Here, $x \in X$ and $u \in U$ are the state and control vectors, $r \in R$ and $y \in Y$ are the measured reference and output, $d \in D$ is the disturbance, $d = T_L$, and $z \in Z$ and $p \in P$ are the unknown and bounded parameter uncertainties.

Our goal is to design the bounded control $u(\cdot)$ within the constrained set

$$U = \{u \in \mathbb{R}^2 : u_{\min} \le u \le u_{\max},\ u_{\min} < 0,\ u_{\max} > 0\} \subset \mathbb{R}^2.$$

An admissible control law, which guarantees a balanced two-phase voltage applied to the ab windings and ensures the maximal electromagnetic torque production, is synthesized as

$$u = \begin{bmatrix} u_{as} \\ u_{bs} \end{bmatrix} = \begin{bmatrix} -\sin(RT\theta_{rm}) & 0 \\ 0 & \cos(RT\theta_{rm}) \end{bmatrix}$$
$$\times \Phi\left(G_x(t)B^T\frac{\partial V(t,x,e)}{\partial x} + G_e(t)B_e^T\frac{\partial V(t,x,e)}{\partial e} + G_i(t)B_e^T\frac{1}{s}\frac{\partial V(t,x,e)}{\partial e}\right)$$

where $e \in E$ is the measured tracking error, $e(t) = r(t) - y(t)$; $\Phi(\cdot)$ is the bounded function (erf, sat, tanh), and $\Phi \in U$, $|\Phi(\cdot)| \le V_{\max}$, $V_{\max}$ is the rated voltage; $G_x(\cdot)$, $G_e(\cdot)$, and $G_i(\cdot)$ are bounded and symmetric, $G_x > 0$, $G_e > 0$, $G_i > 0$; and $V(\cdot)$ is the $C^\kappa(\kappa \ge 1)$ continuously differentiable, real-analytic function.

For $X_0 \subseteq X$, $u \in U$, $r \in R$, $d \in D$, $z \in Z$, and $p \in P$, we obtain the state evolution set X. The state-output set is

$$XY(X_0, U, R, D, Z, P) = \{(x, y) \in X \times Y : x_0 \in X_0,\ u \in U, r \in R, d \in D, z \in Z, p \in P, t \in [t_0, \infty)\}$$

and a *reference-output* map can be found. Our goal is to find the bounded controller such that the tracking error $e(\cdot):[t_0,\infty)\to E$ with $E_0 \subseteq E$ evolves in the specified closed set

$$S_e(\delta) = \{e \in \mathbb{R}^1 : e_0 \in E_0,\ x \in X(X_0, U, R, D, Z, P),\ t \in [t_0,\infty) | \\ \|e(t)\| \le \rho_e(t,\|e_0\|) + \rho_r(\|r\|) + \rho_d(\|d\|) + \rho_y(\|y\|) + \delta,\ \delta \ge 0, \forall e \in E(E_0, R, D, Y), \forall t \in [t_0,\infty)\}$$

Here, $\rho_e(\cdot)$ is the KL-function; $\rho_r(\cdot)$, $\rho_d(\cdot)$ and $\rho_y(\cdot)$ are the K-functions.

A positive-invariant domain of stability is found for the closed-loop system with $x_0 \in X_0$, $e_0 \in E_0$, $u \in U$, $r \in R$, $d \in D$, $z \in Z$ and $p \in P$. In particular,

$$S_s = \{x \in \mathbb{R}^4, e \in \mathbb{R}^1 : \|x(t)\| \le \rho_x(t,\|x_0\|) + \rho_r(\|r\|) + \rho_d(\|d\|) + \delta, \\ \forall x \in X(X_0, U, R, D, Z, P), \forall t \in [t_0,\infty), \|e(t)\| \le \rho_e(t,\|e_0\|) + \rho_r(\|r\|) \\ + \rho_d(\|d\|) + \rho_y(\|y\|) + \delta, \forall e \in E(E_0, R, D, Y), \forall t \in [t_0,\infty)\},$$

where $\rho_x(\cdot)$ is the KL-function.

To study the robustness, tracking, and disturbance rejection, we consider a state-error set

$$XE(X_0, E_0, U, R, D, Z, P) = \{(x, e) \in X \times E : x_0 \in X_0, e_0 \in E_0, u \in U, \\ r \in R, d \in D, z \in Z, p \in P, t \in [t_0,\infty)\}$$

The robust tracking, stability, and disturbance rejection are guaranteed if $XE \subseteq S_s$. The *admissible* set S_s is found by using the Lyapunov stability theory [9], and

$$S_s = \Bigg\{x \in \mathbb{R}^4, e \in \mathbb{R}^1 : x_0 \in X_0, e_0 \in E_0, u \in U, r \in R, d \in D, z \in Z, p \in P | \\ \rho_1\|x\| + \rho_2\|e\| \le V(t,x,e) \le \rho_3\|x\| + \rho_4\|e\|, \frac{dV(t,x,e)}{dt} \le -\rho_5\|x\| - \rho_6\|e\|, \\ \forall x \in X(X_0, U, R, P, Z, P), \forall e \in E(E_0, R, D, Y), \forall t \in [t_0,\infty)\Bigg\}$$

where $\rho_1(\cdot)$, $\rho_2(\cdot)$, $\rho_3(\cdot)$ and $\rho_4(\cdot)$ are the K_∞-functions; and $\rho_5(\cdot)$ and $\rho_6(\cdot)$ are the K-functions.

If in *XE* there exists a C^κ Lyapunov function $V(t, x, \mathrm{e})$ such that for all $x_0 \in X_0, e_0 \in E_0, u \in U$, $r \in R, d \in D, z \in Z$, and $p \in P$ on $[t_0,\infty)$ sufficient condition for stability ($s1$)

$$\rho_1\|x\| + \rho_2\|e\| \le V(t,x,e) \le \rho_3\|x\| + \rho_4\|e\|$$

and inequality

$$\frac{dV(t,x,e)}{dt} \le -\rho_5\|x\| - \rho_6\|e\|$$

which is the sufficient condition for stability $s2$, hold, then

1. solution $x(\cdot):[t_0,\infty)\to X$ for closed-loop system is robustly bounded and stable,
2. convergence of the error vector $e(\cdot):[t_0,\infty)\to E$ to S_e is ensured in *XE*,
3. *XE* is convex and compact, and $XE \subseteq S_s$.

That is, if criteria ($s1$) and ($s2$) are guaranteed, we have $XE \subseteq S_s$.

Using the nonquadratic Lyapunov candidate

$$V(t,x,e)=\sum_{j=0}^{\eta}\frac{2\gamma+1}{2(j+\gamma+1)}\left(x^{\frac{j+\gamma+1}{2\gamma+1}}\right)^T K_{xj}(t)x^{\frac{j+\gamma+1}{2\gamma+1}}+\sum_{j=0}^{\varsigma}\frac{2\beta+1}{2(j+\beta+1)}\left(e^{\frac{j+\beta+1}{2\beta+1}}\right)^T K_{ej}(t)e^{\frac{j+\beta+1}{2\beta+1}}$$
$$+\sum_{i=0}^{\sigma}\frac{2\mu+1}{2(j+\mu+1)}\left(e^{\frac{j+\mu+1}{2\mu+1}}\right)^T K_{ij}(t)e^{\frac{j+\mu+1}{2\mu+1}}$$

one obtains the bounded controller as

$$u=\begin{bmatrix}u_{as}\\u_{bs}\end{bmatrix}=\begin{bmatrix}-\sin(RT\theta_{rm}) & 0\\ 0 & \cos(RT\theta_{rm})\end{bmatrix}\Phi\left(G_x(t)B^T\sum_{j=0}^{\eta}\operatorname{diag}\left[x^{\frac{j-\gamma}{2\gamma+1}}\right]K_{xj}(t)x^{\frac{j+\gamma+1}{2\gamma+1}}\right.$$
$$\left.+G_e(t)B_e^T\sum_{j=0}^{\varsigma}K_{ej}(t)e^{\frac{2j+1}{2\beta+1}}+G_i(t)B_e^T\sum_{j=0}^{\sigma}K_{ij}(t)\frac{1}{s}e^{\frac{2j+1}{2\mu+1}}\right)$$

Here, $K_{xj}(\cdot)$ are the unknown matrix-functions, and $K_{ej}(\cdot)$ and $K_{ij}(\cdot)$ are the unknown coefficients; $\eta = 0, 1, 2,\ldots$; $\gamma = 0, 1, 2,\ldots$; $\varsigma = 0, 1, 2,\ldots$; $\beta = 0, 1, 2,\ldots$; $\sigma = 0, 1, 2,\ldots$; and $\mu = 0, 1, 2,\ldots$.

Under the assumption that X_0, E_0, R, D, Z, and P are admissible, the robust tracking problem is solvable in XE. That is, the bounded real-analytic control $u(\cdot)$ guarantees the robust stability and steers the tracking error to S_e. Furthermore, stability is guaranteed, disturbance rejection is ensured, and specified input-output tracking performance can be achieved.

Applying the controller designed, one maximizes the electromagnetic torque developed by permanent-magnet stepper micromotors. This can be easily shown by using the expression for the electromagnetic torque, the balanced two-phase sinusoidal voltage set (applied phase voltages u_{as} and u_{bs}), as well as the trigonometric identity $\sin^2 a + \cos^2 a = 1$.

The tracking controller can be designed using the tracking error. In particular, we have

$$u=\begin{bmatrix}u_{as}\\u_{bs}\end{bmatrix}=\begin{bmatrix}-\sin(RT\theta_{rm}) & 0\\ 0 & \cos(RT\theta_{rm})\end{bmatrix}\Phi\left(G_e(t)B_e^T\sum_{j=0}^{\varsigma}K_{ei}(t)e^{\frac{2j+1}{2\beta+1}}+G_i(t)B_e^T\sum_{i=0}^{\sigma}K_{ij}(t)\frac{1}{s}e^{\frac{2j+1}{2\mu+1}}\right)$$

The controller design, implementation, and experimental verification are reported in [9].

14.7 Conclusions

This chapter reports the current status, documents innovative results, and researches novel paradigms in synthesis, modeling, analysis, simulation, control, and optimization of high-performance MEMS. These results are obtained applying reported nonlinear modeling, analysis, synthesis, control, and optimization methods which allow one to attain performance assessment and predict outcomes. Novel MEMS were devised. The application of the plate, spherical, torroidal, conical, cylindrical, and asymmetrical motor geometry, as well as *endless*, *open-ended*, and *integrated* electromagnetic systems, allows one to classify MEMS. This idea is extremely useful in the studying of existing MEMS as well as in the synthesis of innovative high-performance MEMS. For example, asymmetrical (unconventional) geometry and *integrated* electromagnetic system can be applied. Optimization can be performed, and the classifier paradigm serves as a starting point from which advanced configurations can be synthesized and straightforwardly interpreted. Microscale motion devices geometry and electromagnetic systems, which play a central role, are related. Structural synthesis and optimization of MEMS are formalized and interpreted using innovative ideas. The MEMS classifier paradigm, in addition to being qualitative, leads one to quantitative analysis. In fact, using the cornerstone laws of electromagnetics and mechanics (e.g., Maxwell's,

Kirchhoff and Newton equations), the differential equations to model electromagnetic and mechanical phenomena and effects can be derived and applied to attain the performance analysis with outcome prediction. Mathematical models for MEMS are found. Making use of these mathematical models, analysis and optimization were performed, and nonlinear control algorithms were designed. The electromagnetics features and phenomena were integrated into the analysis, modeling, synthesis, and optimization. It is shown that to meet the specified level of performance, novel high-performance MEMS should be synthesized, high-fidelity modeling must be performed, advanced controllers have to be synthesized, and highly detailed dynamic nonlinear simulations must be carried out. The results reported have direct application to the analysis and design of high-performance MEMS. Different MEMS can be devised, synthesized, defined, and designed, and a number of long-standing issues related to geometrical variability and electromagnetics are studied. These benchmarking results allow one to reformulate and refine extremely important problems in MEMS theory, and solve a number of very complex issues in design and optimization with the ultimate goal to synthesize innovative high-performance, high torque, and power densities MEMS.

References

1. Lyshevski, S. E., *Nano- and Micro-Electromechanical Systems: Fundamentals of Nano- and Micro-Engineering,* CRC Press, Boca Raton, FL, 2000.
2. Madou, M., *Fundamentals of Microfabrication,* CRC Press, Boca Raton, FL, 1997.
3. Campbell, S. A., *The Science and Engineering of Microelectronic Fabrication,* Oxford University Press, New York, 2001.
4. Lyshevski, S. E., *Electromechanical Systems, Electric Machines, and Applied Mechatronics,* CRC Press, Boca Raton, FL, 1999.
5. Lyshevski, S. E. and Lyshevski, M. A., "Analysis, dynamics, and control of micro-electromechanical systems," *Proc. American Control Conference,* Chicago, IL, pp. 3091–3095, 2000.
6. Mehregany, M. and Tai, Y. C., "Surface micromachined mechanisms and micro-motors," *J. Micromechanics and Microengineering,* vol. 1, pp. 73–85, 1992.
7. Becker, E. W., Ehrfeld, W., Hagmann, P., Maner, A., and Mynchmeyer, D., "Fabrication of microstructures with high aspect ratios and great structural heights by synchrotron radiation lithography, galvanoformung, and plastic molding (LIGA process)," *Microelectronic Engineering,* vol. 4, pp. 35–56, 1986.
8. Guckel, H., Christenson, T. R., Skrobis, K. J., Klein, J., and Karnowsky, M., "Design and testing of planar magnetic micromotors fabricated by deep X-ray lithography and electroplating," *Technical Digest of International Conference on Solid-State Sensors and Actuators, Transducers 93,* Yokohama, Japan, pp. 60–64, 1993.
9. Lyshevski, S. E., *Control Systems Theory with Engineering Applications,* Birkhäuser, Boston, MA, 2001.

15

The Physical Basis of Analogies in Physical System Models

Neville Hogan
Massachusetts Institute of Technology

Peter C. Breedveld
University of Twente

15.1 Introduction

One of the fascinating aspects of mechatronic systems is that their function depends on interactions between electrical and mechanical behavior and often magnetic, fluid, thermal, chemical, or other effects as well. At the same time, this can present a challenge as these phenomena are normally associated with different disciplines of engineering and physics. One useful approach to this multidisciplinary or "multi-physics" problem is to establish analogies between behavior in different domains—for example, resonance due to interaction between inertia and elasticity in a mechanical system is analogous to resonance due to interaction between capacitance and inductance in an electrical circuit. Analogies can provide valuable insight about how a design works, identify equivalent ways a particular function might be achieved, and facilitate detailed quantitative analysis. They are especially useful in studying dynamic behavior, which often arises from interactions between domains; for example, even in the absence of elastic effects, a mass moving in a magnetic field may exhibit resonant oscillation. However, there are many ways that analogies may be established and, unfortunately, the most appropriate analogy between electrical circuits, mechanical and fluid systems remains unresolved: is force like current, or is force more like voltage? In this contribution we examine the physical basis of the analogies in common use and how they may be extended beyond mechanical and electrical systems.

15.2 History

It is curious that one of the earliest applications of analogies between electrical and mechanical systems was to enable the demonstration and study of transients in electrical networks that were otherwise too fast to be observed by the instrumentation of the day by identifying mechanical systems with equivalent dynamic behavior; that was the topic of a series of articles on "Models and analogies for demonstrating electrical principles" (*The Engineer,* 1926). Improved methods capable of observing fast electrical transients directly (especially the cathode ray oscilloscope, still in use today) rendered this approach obsolete but enabled quantitative study of nonelectrical systems via analogous electrical circuits (Nickle, 1925). Although that method had considerably more practical importance at the time than it has today (we now have the luxury of vastly more powerful tools for numerical computation of electromechanical system responses), in the late '20s and early '30s a series of papers (Darrieus, 1929; Hähnle, 1932; Firestone, 1933) formulated a rational method to use electrical networks as a framework for establishing analogies between physical systems.

15.3 The Force-Current Analogy: Across and Through Variables

Firestone identified two types of variable in each physical domain—"across" and "through" variables—which could be distinguished based on how they were measured. An "across" variable may be measured as a difference between values at two points in space (conceptually, *across* two points); a "through" variable may be measured by a sensor in the path of power transmission between two points in space (conceptually, it is transmitted *through* the sensor). By this classification, electrical voltage is analogous to mechanical velocity and electrical current is analogous to mechanical force. Of course, this classification of variables implies a classification of network elements: a mass is analogous to a capacitor, a spring is analogous to an inductor and so forth.

The "force-is-like-current" or "mass-capacitor" analogy has a sound mathematical foundation. Kirchhoff's node law or current law, introduced in 1847 (the sum of currents into a circuit node is identically zero) can be seen as formally analogous to D'Alembert's principle, introduced in 1742 (the sum of forces on a body is identically zero, provided the sum includes the so-called "inertia force," the negative of the body mass times its acceleration). It is the analogy used in linear-graph representations of lumped-parameter systems, proposed by Trent in 1955. Linear graphs bring powerful results from mathematical graph theory to bear on the analysis of lumped-parameter systems. For example, there is a systematic procedure based on partitioning a graph into its *tree* and *links* for selecting sets of independent variables to describe a system. Graph-theoretic approaches are closely related to matrix methods that in turn facilitate computer-aided methods. Linear graphs provide a unified representation of lumped-parameter dynamic behavior in several domains that has been expounded in a number of successful textbooks (e.g., Shearer et al., 1967; Rowell & Wormley, 1997).

The mass-capacitor analogy also appears to afford some practical convenience. It is generally easier to identify points of common velocity in a mechanical system than to identify which elements experience the same force; and it is correspondingly easier to identify the nodes in an electrical circuit than all of its loops. Hence with this analogy it is straightforward to identify an electrical network equivalent to a mechanical system, at least in the one-dimensional case.

Drawbacks of the Across-Through Classification

Despite the obvious appeal of establishing analogies based on practical measurement procedures, the force-current analogy has some drawbacks that will be reviewed below: (i) on closer examination, measurement-based classification is ambiguous; (ii) its extension to more than one-dimensional mechanical systems is problematical; and (iii) perhaps most important, it leads to analogies (especially between mechanical and fluid systems) that defy common physical insight.

Measurement as a Basis for Analogies

Even a cursory review of state-of-the-art measurement technologies shows that the across-through classification may be an anachronism or, at best, an over-simplification. Velocity (an "across" variable) may be measured by an integrating accelerometer that is attached only to the point where velocity is measured—that's how the human inner ear measures head velocity. While the velocity is measured with respect to an inertial reference frame (as it should be), there is no tangible connection to that frame. As a further example, current in a conductor (a "through" variable) may be measured without inserting an ammeter in the current path; sensors that measure current by responding to the magnetic field next to the conductor are commercially available (and preferred in some applications). Moreover, in some cases similar methods can be applied to measure both "across" and "through" variables. For example, fluid flow rate is classified as a through variable, presumably by reference to its measurement by, for example, a positive-displacement meter in the flow conduit; that's the kind of fluid measurement commonly used in a household water meter. However, optical methods that are used to measure the velocity of a rigid body (classified as an across variable) are often adapted to measure the volumetric flow rate of a fluid (laser doppler velocimetry is a notable example). Apparently the same fundamental measurement technology can be associated with an across variable in one domain and a through variable in another. Thus, on closer inspection, the definition of across and through variables based on measurement procedures is, at best, ambiguous.

Beyond One-Dimensional Mechanical Systems

The apparent convenience of equating velocities in a mechanical system with voltages at circuit nodes diminishes rapidly as we go beyond translation in one dimension or rotation about a fixed axis. A translating body may have two or three independent velocities (in planar and spatial motion, respectively). Each independent velocity would appear to require a separate independent circuit node, but the kinetic energy associated with translation can be redistributed at will among these two or three degrees of freedom (e.g., during motion in a circle at constant speed the total kinetic energy remains constant while that associated with each degree of freedom varies). This requires some form of connection between the corresponding circuit nodes in an equivalent electrical network, but what that connection should be is not obvious.

The problem is further exacerbated when we consider rotation. Even the simple case of planar motion (i.e., a body that may rotate while translating) requires three independent velocities, hence three independent nodes in an equivalent electrical network. Reasoning as above we see that these three nodes must be connected but in a different manner from the connection between three nodes equivalent to spatial translation. Again, this connection is hardly obvious, yet translating while rotating is ubiquitous in mechanical systems—that's what a wheel usually does.

Full spatial rotation is still more daunting. In this case interaction between the independent degrees of freedom is especially important as it gives rise to gyroscopic effects, including oscillatory precession and nutation. These phenomena are important practical considerations in modern mechatronics, not arcane subtleties of classical mechanics; for example, they are the fundamental physics underlying several designs for a microelectromechanical (MEMS) vibratory rate gyroscope (Yazdi et al., 1998).

Physical Intuition

In our view the most important drawback of the across-through classification is that it identifies force as analogous to fluid flow rate as well as electrical current (with velocity analogous to fluid pressure as well as voltage). This is highly counter-intuitive and quite confusing. By this analogy, fluid pressure is not analogous to force despite the fact that pressure is commonly defined as force per unit area. Furthermore, stored kinetic energy due to fluid motion is not analogous to stored kinetic energy due to motion of a rigid body. Given the remarkable similarity of the physical processes underlying these two forms of energy storage, it is hard to understand why they should not be analogous.

Insight is the ultimate goal of modeling. It is a crucial factor in producing innovative and effective designs and depends on developing and maintaining a "physical intuition" about the way devices behave. It is important that analogies between physical effects in different domains can be reconciled with the physical intuition and any method that requires a counter-intuitive analogy is questionable; at a minimum it warrants careful consideration.

15.4 Maxwell's Force-Voltage Analogy: Effort and Flow Variables

An alternative analogy classifies variables in each physical domain that (loosely speaking) describe motion or cause it. Thus fluid flow rate, electrical current, and velocity are considered analogous (sometimes generically described as "flow" variables). Conversely, fluid pressure, electrical voltage, and force are considered analogous (sometimes generically described as "effort" variables).

The "force-is-like-voltage" analogy is the oldest drawn between mechanical and electrical systems. It was first proposed by Maxwell (1873) in his treatise on electricity and magnetism, where he observed the similarity between the Lagrangian equations of classical mechanics and electromechanics. That was why Firestone (1933) presented his perspective that force is like current as "A *new* analogy between mechanical and electrical systems" (emphasis added). Probably because of its age, the force-voltage analogy is deeply embedded in our language. In fact, voltage is still referred to as "electromotive force" in some contexts. Words like "resist" or "impede" also have this connotation: a large resistance or impedance implies a large force for a given motion or a large voltage for a given current.

In fact, Maxwell's classification of velocity as analogous to electrical current (with force analogous to voltage) has a deeper justification than the similarity of one mathematical form of the equations of mechanics and electromechanics; it can be traced to a similarity of the underlying physical processes.

Systems of Particles

Our models of the physical world are commonly introduced by describing systems of particles distributed in space. The particles may have properties such as mass, charge, etc., though in a given context we will deliberately choose to neglect most of those properties so that we may concentrate on a single physical phenomenon of interest. Thus, to describe electrical capacitance, we consider only charge, while to describe translational inertia, we consider only mass and so forth.

Given that this common conceptual model is used in different domains, it may be used to draw analogies between the variables of different physical domains. From this perspective, quantities associated with the motion of particles may be considered analogous to one another; thus mechanical velocity, electrical current, and fluid flow rate are analogous. Accordingly, mechanical displacement, displaced fluid volume, and displaced charge are analogous; and thus force, fluid pressure, and voltage are analogous. This classification of variables obviously implies a classification of network elements: a spring relates mechanical displacement and force; a capacitor relates displaced charge and voltage. Thus a spring is analogous to a capacitor, a mass to an inductor, and for this reason, this analogy is sometimes termed the "mass-inductor" analogy.

Physical Intuition

The "system-of-particles" models naturally lead to the "intuitive" analogy between pressure, force, and voltage. But, is such a vague and ill-defined concept as "physical intuition" an appropriate consideration in drawing analogies between physical systems? After all, physical intuition might largely be a matter of usage and familiarity, rooted in early educational and cultural background.

We think not; instead we speculate that physical intuition may be related to conformity with a mental model of the physical world. That mental model is important for thinking about physical systems and, if shared, for communicating about them. Because the "system-of-particles" model is widely assumed

(sometimes explicitly, sometimes implicitly) in the textbooks and handbooks of basic science and engineering we speculate that it may account for the physical intuition shared by most engineers. If so, then conforming with that common "system-of-particles" mental model is important to facilitate designing, thinking, and communicating about mechatronic systems. The force-voltage analogy does so; the force-current analogy does not.

Dependence on Reference Frames

The "system-of-particles" model also leads to another important physical consideration in the choice of analogies between variables: the way they depend on reference frames. The mechanical displacement that determines the elastic potential energy stored in a spring and the displaced charge that determines the electrostatic potential energy stored in a capacitor may be defined with respect to *any* reference frame (whether time-varying or stationary). In contrast, the motion required for kinetic energy storage in a rigid body or a fluid must be defined with respect to an *inertial* frame. Though it may often be overlooked, the motion of charges required for magnetic field storage must also be defined with respect to an inertial frame (Feynman et al., 1963).

To be more precise, the constitutive equations of energy storage based on motion (e.g., in a mass or an inductor) require an inertial reference frame (or must be modified in a non-inertial reference frame). In contrast, the constitutive equations of energy storage based on displacement (e.g., in a spring or a capacitor) do not. Therefore, the mass-inductor (force-voltage) analogy is more consistent with fundamental physics than the mass-capacitor (force-current) analogy.

The modification of the constitutive equations for magnetic energy storage in a non-inertial reference frame is related to the transmission of electromagnetic radiation. However, Kirchhoff's laws (more aptly termed "Kirchhoff's approximations"), which are the foundations of electric network theory, are equivalent to assuming that electromagnetic radiation is absent or negligible. It might, therefore, be argued that the dependence of magnetic energy storage on an inertial reference frame is negligible for electrical circuits, and hence is irrelevant for any discussion of the physical basis of analogies between electrical circuits and other lumped-parameter dynamic-system models. That is undeniably true and could be used to justify the force-current analogy. Nevertheless, because of the confusion that can ensue, the value of an analogy that is fundamentally inconsistent with the underlying physics of lumped-parameter models is questionable.

15.5 A Thermodynamic Basis for Analogies

Often in the design and analysis of mechatronic systems it is necessary to consider a broader suite of phenomena than those of mechanics and electromechanics. For instance, it may be important to consider thermal conduction, convection, or even chemical reactions and more. To draw analogies between the variables of these domains it is helpful to examine the underlying physics. The analogous dynamic behavior observed in different physical domains (resonant oscillation, relaxation to equilibrium, etc.) is not merely a similarity of *mathematical* forms, it has a common *physical* basis which lies in the storage, transmission, and irreversible dissipation of energy. Consideration of energy leads us to thermodynamics; we show next that thermodynamics provides a broader basis for drawing analogies and yields some additional insight.

All of the displacements considered to be analogous above (i.e., mechanical displacement, displaced fluid volume, and displaced charge) may be associated with an energy storage function that requires equilibrium for its definition, the displacement being the argument of that energy function. Generically, these may be termed potential energy functions. To elaborate, elastic energy storage requires sustained but recoverable deformation of a material (e.g., as in a spring); the force required to sustain that deformation is determined at equilibrium, defined when the time rate of change of relative displacement of the material particles is uniformly zero (i.e., all the particles are at rest relative to each other). Electrostatic energy storage requires sustained separation of mobile charges of opposite sign (e.g., as in

a capacitor); the required voltage is determined at equilibrium, defined when the time rate of change of charge motion is zero (i.e., all the charges are at rest relative to each other).

Extensive and Intensive Variables

In the formalism of thermodynamics, the amount of stored energy and the displacement that determines it are *extensive* variables. That is, they vary with the spatial extent (i.e., size or volume) of the object storing the energy. The total elastic energy stored in a uniform rod of constant cross-sectional area in an idealized uniform state of stress is proportional to the length (and hence volume) of the rod; so is the total relative displacement of its ends; both are extensive variables. The total electrostatic energy stored in an idealized parallel-plate capacitor (i.e., one with no fringe fields) is proportional to the area of the plates (and hence, for constant gap, the volume they enclose); so is the total separated charge on the plates; both are extensive variables (cf., Breedveld, 1984).

Equilibrium of these storage elements is established by an *intensive* variable that does not change with the size of the object. This variable is the gradient (partial derivative) of the stored energy with respect to the corresponding displacement. Thus, at equilibrium, the force on each cross-section of the rod is the same regardless of the length or volume of the rod; force is an intensive variable. If the total charge separated in the capacitor is proportional to area, the voltage across the plates is independent of area; voltage is an intensive variable.

Dynamics is not solely due to the storage of energy but arises from the transmission and deployment of power. The instantaneous power into an equilibrium storage element is the product of the (intensive) gradient variable (force, voltage) with the time rate of change of the (extensive) displacement variable (velocity, current). Using this thermodynamics-based approach, all intensive variables are considered analogous, as are all extensive variables and their time rates of change, and so on.

Drawing analogies from a thermodynamic classification into extensive and intensive variables may readily be applied to fluid systems. Consider the potential energy stored in an open container of incompressible fluid: The pressure at any specified depth is independent of the area at that depth and the volume of fluid above it; pressure is an intensive variable analogous to force and voltage, as our common physical intuition suggests it should be. Conversely, the energy stored in the fluid above that depth is determined by the volume of fluid; energy and volume are extensive variables, volume playing the role of displacement analogous to electrical charge and mechanical displacement. Pressure is the partial derivative of stored energy with respect to volume and the instantaneous power into storage is the product of pressure with volumetric flow rate, the time rate of change of volume flowing past the specified depth.

An important advantage of drawing analogies from a classification into extensive and intensive variables is that it may readily be generalized to domains to which the "system-of-particles" image may be less applicable. For example, most mechatronic designs require careful consideration of heating and cooling but there is no obvious flow of particles associated with heat flux. Nevertheless, extensive and intensive variables associated with equilibrium thermal energy storage can readily be identified. Drawing on classical thermodynamics, it can be seen that (total) entropy is an extensive variable and plays the role of a displacement. The gradient of energy with respect to energy is temperature, an intensive variable, which should be considered analogous to force, voltage, and pressure. Equality of temperature establishes thermal equilibrium between two bodies that may store heat (energy) and communicate it to one another.

A word of caution is appropriate here as a classification into extensive and intensive variables properly applies only to scalar quantities such as pressure, volume, etc. As outlined below, the classification can be generalized in a rigorous way to nonscalar quantities, but care is required (cf., Breedveld, 1984).

Equilibrium and Steady State

In some (though not all) domains energy storage may also be based on motion. Kinetic energy storage may be associated with rigid body motion or fluid motion; magnetic energy storage requires motion of charges. The thermodynamics-based classification properly groups these different kinds of energy storage as analogous to one another and generically they may be termed *kinetic* energy storage elements.

All of the motion variables considered to be analogous (i.e., velocity, fluid flow rate, current) may be associated with an energy storage function that is defined by *steady state* (rather than by equilibrium). For a rigid body, steady motion requires zero net force, and hence constant momentum and kinetic energy. For the magnetic field that stores energy in an inductor, steady current requires zero voltage, and hence constant magnetic flux and magnetic energy.

It might reasonably be argued that any distinction between equilibrium and steady state is purely a matter of perspective and common usage, rather than a fundamental feature of the physical world. For example, with an alternative choice of reference frames, "sustained motion" could be redefined as "rest" or "equilibrium." From this perspective, a zero-relative-velocity "equilibrium" between two rigid bodies (or between a rigid body and a reference frame) could be defined by zero force. Following this line of reasoning any distinction between the mass-inductor and mass-capacitor analogies would appear to be purely a matter of personal choice. However, while the apparent equivalence of "equilibrium" and "steady state" may be justifiable in the formal mathematical sense of zero rate of change of a variable, in a mechanical system, displacement (or position) and velocity (or momentum) are fundamentally different. For example, whereas velocity, force, and momentum may be transformed between reference frames as rank-one tensors, position (or displacement) may not be transformed as a tensor of any kind. Thus, a distinction between equilibrium and steady state reflects an important aspect of the structure of physical system models.

Analogies, Not Identities

It is important to remember that any classification to establish analogies is an abstraction. At most, dynamic behavior in different domains may be similar; it is not identical. We have pointed out above that if velocity or current is used as the argument of an energy storage function, care must be taken to identify an appropriate inertial reference frame and/or to understand the consequences of using a non-inertial frame. However, another important feature of these variables is that they are fundamentally vectors (i.e., they have a definable spatial orientation). One consequence is that the thermodynamic definition of extensive and intensive variables must be generalized before it may be used to classify these variables (cf., Breedveld, 1984). In contrast, a quantity such as temperature or pressure is fundamentally a scalar. Furthermore, both of these quantities are intrinsically "positive" scalars insofar as they have well-defined, unique and physically meaningful zero values (absolute zero temperature, the pressure of a perfect vacuum). Quite aside from any dependence on inertial reference frames, the across-through analogy between velocity (a vector with no unique zero value) and pressure (a scalar with a physically important zero) will cause error and confusion if used without due care.

This consideration becomes especially important when similar elements of a model are combined (for example, a number of bodies moving with identical velocity may be treated as a single rigid body) to simplify the expression of dynamic equations or improve their computability. The engineering variables used to describe energy storage can be categorized into two groups: (i) positive-valued scalar variables and (ii) nonscalar variables. Positive-valued scalar variables have a physically meaningful zero or absolute reference; examples include the volume of stored fluid, the number of moles of a chemical species, entropy, etc. Nonscalar[1] variables have a definable spatial orientation. Even in the one-dimensional case they can be positive or negative, the sign denoting direction with respect to some reference frame; examples include displacement, momentum, etc. These variables generally do not have a physically meaningful zero or absolute reference, though some of them must be defined with respect to an inertial frame.

Elements of a model that describe energy storage based on scalar variables can be combined in only one way: they must be in mutual equilibrium; their extensive variables are added, while the corresponding intensive variables are equal, independent of direction, and determine the equilibrium condition. For model elements that describe energy storage based on nonscalar variables there are usually two options.

[1]The term "vector variables" suggests itself but these variables may include three-dimensional spatial orientation, which may not be described as a vector.

Electrical capacitors, for instance, may be combined in parallel or in series and the resulting equivalent capacitor may readily be determined. In a parallel connection, equilibrium is determined by voltage (an intensive variable) and the electric charges (extensive variables) are added as before. However, a series connection is the "dual" in the sense that the roles of charge and voltage are exchanged: equality of charges determines equilibrium and the voltages are added. Mechanical springs may also be combined in two ways. However, that is not the case for translational masses and rotational inertias; they may only be combined into a single equivalent rigid body if their velocities are equal and in that case their momenta are added.

The existence of two "dual" ways to combine some, but not all, of the energy storage elements based on nonscalar quantities is somewhat confusing. It may have contributed to the lengthy debate (if we date its beginning to Maxwell, lasting for over a century!) on the best analogy between mechanical and electrical systems. Nevertheless, the important point is that series and parallel connections may not be generalized in a straightforward way to all domains.

Nodicity

As insight is the foremost goal of modeling, analogies should be chosen to promote insight. Because there may be fundamental differences between all of the physical domains, care should be exercised in drawing analogies to ensure that special properties of one domain should not be applied inappropriately to other domains. This brings us to what may well be the strongest argument against the across-through classification. History suggests that it originated with the use of equivalent electrical network representations of nonelectrical systems. Unfortunately, electrical networks provide an inappropriate basis for developing a general representation of physical system dynamics. This is because electrical networks enjoy a special property, *nodicity*, which is quite unusual among the physical system domains (except as an approximation).

Nodicity refers to the fact that any sub-network (cut-set) of an electrical network behaves as a node in the sense that a Kirchhoff current balance equation may be written for the entire sub-network. As a result of nodicity, electrical network elements can be assembled in arbitrary topologies and yet still describe a physically realizable electrical network. This property of "arbitrary connectability" is not a general property of lumped-parameter physical system models. Most notably, mass elements cannot be connected arbitrarily; they must always be referenced to an inertial frame. For that reason, electrical networks can be quite misleading when used as a basis for a general representation of physical system dynamics. This is not merely a mathematical nicety; some consequences of non-nodic behavior for control system analysis have recently been explored (Won and Hogan, 1998).

By extension, because each of the physical domains has its unique characteristics, any attempt to formulate analogies by taking one of the domains (electrical, mechanical, or otherwise) as a starting point is likely to have limitations. A more productive approach is to begin with those characteristics of physical variables common to all domains and that is the reason to turn to thermodynamics. In other words, the best way to identify analogies *between* domains may be to "step outside" *all* of them. By design, general characteristics of all domains such as the extensive nature of stored energy, the intensive nature of the variables that define equilibrium, and so forth, are not subject to the limitations of any one (such as nodicity). That is the main advantage of drawing analogies based on thermodynamic concepts such as the distinction between extensive and intensive variables.

15.6 Graphical Representations

Analogies are often associated with abstract graphical representations of multi-domain physical system models. The force-current analogy is usually associated with the linear graph representation of networks introduced by Trent (1955); the force-voltage analogy is usually associated with the bond graph representation introduced by Paynter (1960). Bond graphs classify variables into efforts (commonly force, voltage, pressure, and so forth) and flows (commonly velocity, current, fluid flow rate, and so forth). Bond graphs extend all the practical benefits of the force-current (across-through) analogy to the force-voltage (effort-flow) analogy: they provide a unified representation of lumped-parameter dynamic behavior in several

domains that has been expounded in a number of successful textbooks (e.g., Karnopp et al., 1975, 1999), there are systematic methods for selecting sets of independent variables to describe a system, ways to take advantage of the ease of identifying velocities and voltages, and matrix methods to facilitate computer analysis. In fact, several computer-aided modeling support packages using the bond-graph language are now available. Furthermore, bond graphs have been applied successfully to describe the dynamics of spatial mechanisms (including gyroscopic effects) while, to the authors' knowledge, linear graphs have not.

Although the force-voltage analogy is most commonly used with bond graphs, the force-current analogy can be used just as readily; the underlying mathematical formalism is indifferent to the choice of which variables are chosen as analogous. In fact, pursuing this line of thought, the choice is unnecessary and may be avoided; doing so affords a way to clarify the potential confusion over the role of intensive variables and the dual types of connection available for some elements in some domains.

In the Generalized Bond Graph (GBG) approach (Breedveld, 1984) all energy storage becomes analogous and only one type of storage element, a (generalized) capacitor, is identified. Its displacement is an extensive variable; the gradient of its energy storage function with respect to that displacement is an intensive variable. In some (but not all) domains a particular kind of coupling known as a *gyrator* is found that gives rise to the *appearance* of a dual type of energy storage, a (generalized) inertia as well as the possibility of dual ways to connect elements. The GBG representation emphasizes the point that the presence of dual types of energy storage and dual types of connection is a special property (albeit an important one) of a limited number of domains. In principle, either a "mass-capacitor" analogy or a "mass-inductor" analogy can be derived from a GBG representation by choosing to associate the gyrating coupling with either the "equilibrium" or "steady-state" energy storage elements.

The important point to be taken here is that the basis of analogies between domains does not depend on the use of a particular abstract graphical representation. The practical value of establishing analogies between domains and the merits of a domain-independent approach based on intensive vs. extensive variables remains regardless of which graph-theoretic tools (if any) are used for analysis.

15.7 Concluding Remarks

In the foregoing we articulated some important considerations in the choice of analogies between variables in different physical domains. From a strictly mathematical viewpoint there is little to choose; both analogies may be used as a basis for rigorous, self-consistent descriptions of physical systems. The substantive and important factors emerge from a physical viewpoint—considering the structured way physical behavior is described in the different domains. Summarizing:

- The "system-of-particles" model that is widely assumed in basic science and engineering naturally leads to the intuitive analogy between force and voltage, velocity and current, a mass and an inductor, and so on.
- The measurement procedures used to motivate the distinction between across and through variables at best yield an ambiguous classification.
- Nodicity (the property of "arbitrary connectability") is not a general property of lumped-parameter physical system models. Thus, electrical networks, which are nodic, can be quite misleading when used as a basis for a general representation of physical system dynamics.
- The intuitive analogy between velocity and current is consistent with a thermodynamic classification into extensive and intensive variables. As a result, the analogy can be generalized to dynamic behavior in domains to which the "system-of-particles" image may be less applicable.
- The force-voltage or mass-inductor analogy reflects an important distinction between equilibrium energy-storage phenomena and steady-state energy-storage phenomena: the constitutive equations of steady-state energy storage phenomena require an inertial reference frame (or must be modified in a non-inertial reference frame) while the constitutive equations of equilibrium energy storage phenomena do not.

Our reasoning is based on an assumption that models of physical system dynamics should properly reflect the way descriptions of physical phenomena depend on reference frames and should be compatible with thermodynamics. The across-through classification of variables does not meet these requirements. By contrast, the classification of variables based on the system-of-particles point of view that leads to an analogy between force, pressure, and voltage on the one hand and velocity, fluid flow, and current on the other not only satisfies these criteria, but is the least artificial from a common-sense point of view. We believe this facilitates communication and promotes insight, which are the ultimate benefits of using analogies.

Acknowledgments

Neville Hogan was supported in part by grant number AR40029 from the National Institutes of Health.

References

(1926). Models and analogies for demonstrating electrical principles, parts I-XIX. *The Engineer,* 142.

Breedveld, P.C. (1984). *Physical Systems Theory in Terms of Bond Graphs,* University of Twente, Enschede, Netherlands, ISBN 90-9000599-4 (distr. by author).

Darrieus, M. (1929). Les modeles mecaniques en electrotechnique. Leur application aux problemes de stabilite. *Bull. Soc. Franc. Electric.,* 36:729–809.

Feynman, R.P., Leighton, R.B., and Sands, M. (1963). *The Feynman Lectures on Physics, Volume II: Mainly Electromagnetism and Matter,* Addison-Wesley Publishing Company.

Firestone, F.A. (1933). A new analogy between mechanical and electrical system elements. *Journal of the Acoustic Society of America,* 3:249–267.

Hähnle, W. (1932). Die darstellung elektromechanischer gebilde durch rein elektrsiche schaltbilder. *Wissenschaftliche Veroffentl. Siemens Konzern,* 11:1–23.

Karnopp, D.C. and Rosenberg, R.C. (1975). *System Dynamics: A Unified Approach,* John Wiley.

Karnopp, D.C., Margolis, D.L., and Rosenberg, R.C. (1999). *System Dynamics: Modeling and Simulation of Mechatronic Systems,* 3rd edition, John Wiley.

Maxwell, J.C. (1873). *Treatise on Electricity and Magnetism*.

Nickle, C.A. (1925). Oscillographic solutions of electro-mechanical systems. *Trans. A.I.E.E.,* 44:844–856.

Rowell, D. and Wormley, D.N. (1997). *System Dynamics: An Introduction,* Prentice-Hall, NJ.

Shearer, J.L., Murphy, A.T., and Richardson, H.H. (1967). *Introduction to System Dynamics,* Addison-Wesley Publishing Company.

Trent, H.M. (1955). Isomorphisms between oriented linear graphs and lumped physical systems. *Journal of the Acoustic Society of America,* 27:500–527.

Won, J. and Hogan, N. (1998). Coupled stability of non-nodic physical systems. *IFAC Symposium on Nonlinear Control Systems Design.*

Yazdi, N., Ayazi, F., and Najafi, K. (1998). Micromachined inertial sensors. *Proc. IEEE,* 86(8), 1640–1659.

Sensors and Actuators

16

Introduction to Sensors and Actuators

M. Anjanappa
University of Maryland Baltimore County

K. Datta
University of Maryland Baltimore County

T. Song
University of Maryland Baltimore County

Sensors and actuators are two critical components of every closed loop control system. Such a system is also called a *mechatronics system*. A typical mechatronics system as shown in Fig. 16.1 consists of a sensing unit, a controller, and an actuating unit. A sensing unit can be as simple as a single sensor or can consist of additional components such as filters, amplifiers, modulators, and other signal conditioners. The controller accepts the information from the sensing unit, makes decisions based on the control algorithm, and outputs commands to the actuating unit. The actuating unit consists of an actuator and optionally a power supply and a coupling mechanism.

16.1 Sensors

Sensor is a device that when exposed to a physical phenomenon (temperature, displacement, force, etc.) produces a proportional output signal (electrical, mechanical, magnetic, etc.). The term transducer is often used synonymously with sensors. However, ideally, a sensor is a device that responds to a change in the physical phenomenon. On the other hand, a transducer is a device that converts one form of energy into another form of energy. Sensors are transducers when they sense one form of energy input and output in a different form of energy. For example, a thermocouple responds to a temperature change (thermal energy) and outputs a proportional change in electromotive force (electrical energy). Therefore, a thermocouple can be called a sensor and or transducer.

Classification

Table 16.1 lists various types of sensors that are classified by their measurement objectives. Although this list is by no means exhaustive, it covers all the basic types including the new generation sensors such as smart material sensors, microsensors, and nanosensors.

TABLE 16.1 Type of Sensors for Various Measurement Objectives

Sensor	Features
	Linear/Rotational sensors
Linear/Rotational variable differential transducer (LVDT/RVDT)	High resolution with wide range capability
	Very stable in static and quasi-static applications
Optical encoder	Simple, reliable, and low-cost solution
	Good for both absolute and incremental measurements
Electrical tachometer	Resolution depends on type such as generator or magnetic pickups
Hall effect sensor	High accuracy over a small to medium range
Capacitive transducer	Very high resolution with high sensitivity
	Low power requirements
	Good for high frequency dynamic measurements
Strain gauge elements	Very high accuracy in small ranges
	Provides high resolution at low noise levels
Interferometer	Laser systems provide extremely high resolution in large ranges
	Very reliable and expensive
Magnetic pickup	Output is sinusoidal
Gyroscope	
Inductosyn	Very high resolution over small ranges
	Acceleration sensors
Seismic accelerometer	Good for measuring frequencies up to 40% of its natural frequency
Piezoelectric accelerometer	High sensitivity, compact, and rugged
	Very high natural frequency (100 kHz typical)
	Force, torque, and pressure sensor
Strain gauge	Good for both static and dynamic measurements
Dynamometers/load cells	They are also available as micro- and nanosensors
Piezoelectric load cells	Good for high precision dynamic force measurements
Tactile sensor	Compact, has wide dynamic range, and high
Ultrasonic stress sensor	Good for small force measurements
	Flow sensors
Pitot tube	Widely used as a flow rate sensor to determine speed in aircrafts
Orifice plate	Least expensive with limited range
Flow nozzle, venturi tubes	Accurate on wide range of flow
	More complex and expensive
Rotameter	Good for upstream flow measurements
	Used in conjunction with variable inductance sensor
Ultrasonic type	Good for very high flow rates
	Can be used for both upstream and downstream flow measurements
Turbine flow meter	Not suited for fluids containing abrasive particles
	Relationship between flow rate and angular velocity is linear
Electromagnetic flow meter	Least intrusive as it is noncontact type
	Can be used with fluids that are corrosive, contaminated, etc.
	The fluid has to be electrically conductive
	Temperature sensors
Thermocouples	This is the cheapest and the most versatile sensor
	Applicable over wide temperature ranges (-200∞C to 1200∞C typical)
Thermistors	Very high sensitivity in medium ranges (up to 100∞C typical)
	Compact but nonlinear in nature
Thermodiodes, thermo transistors	Ideally suited for chip temperature measurements
	Minimized self heating
RTD—resistance temperature detector	More stable over a long period of time compared to thermocouple
	Linear over a wide range

(*continued*)

TABLE 16.1 Type of Sensors for Various Measurement Objectives (Continued)

Sensor	Features
Infrared type	Noncontact point sensor with resolution limited by wavelength
Infrared thermography	Measures whole-field temperature distribution
	Proximity sensors
Inductance, eddy current, hall effect, photoelectric, capacitance, etc.	Robust noncontact switching action The digital outputs are often directly fed to the digital controller
	Light sensors
Photoresistors, photodiodes, photo transistors, photo conductors, etc.	Measure light intensity with high sensitivity Inexpensive, reliable, and noncontact sensor
Charge-coupled diode	Captures digital image of a field of vision
	Smart material sensors
Optical fiber	
As strain sensor	Alternate to strain gages with very high accuracy and bandwidth Sensitive to the reflecting surface's orientation and status
As level sensor	Reliable and accurate
As force sensor	High resolution in wide ranges
As temperature sensor	High resolution and range (up to 2000∞C)
Piezoelectric	
As strain sensor	Distributed sensing with high resolution and bandwidth
As force sensor	Most suitable for dynamic applications
As accelerometer	Least hysteresis and good setpoint accuracy
Magnetostrictive	
As force sensors	Compact force sensor with high resolution and bandwidth Good for distributed and noncontact sensing applications
As torque sensor	Accurate, high bandwidth, and noncontact sensor
	Micro- and nano-sensors
Micro CCD image sensor	Small size, full field image sensor
Fiberscope	Small (0.2 mm diameter) field vision scope using SMA coil actuators
Micro-ultrasonic sensor	Detects flaws in small pipes
Micro-tactile sensor	Detects proximity between the end of catheter and blood vessels

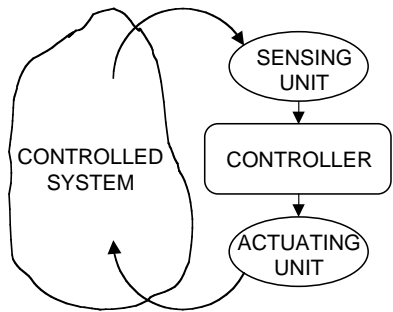



FIGURE 16.1 A typical mechatronics system.

Sensors can also be classified as *passive* or *active*. In passive sensors, the power required to produce the output is provided by the sensed physical phenomenon itself (such as a thermometer) whereas the active sensors require external power source (such as a strain gage).

Furthermore, sensors are classified as *analog* or *digital* based on the type of output signal. Analog sensors produce continuous signals that are proportional to the sensed parameter and typically require

analog-to-digital conversion before feeding to the digital controller. Digital sensors on the other hand produce digital outputs that can be directly interfaced with the digital controller. Often, the digital outputs are produced by adding an analog-to-digital converter to the sensing unit. If many sensors are required, it is more economical to choose simple analog sensors and interface them to the digital controller equipped with a multi-channel analog-to-digital converter.

Principle of Operation

Linear and Rotational Sensors

Linear and rotational position sensors are two of the most fundamental of all measurements used in a typical mechatronics system. The most common type position sensors are listed in Table 16.1. In general, the position sensors produce an electrical output that is proportional to the displacement they experience. There are contact type sensors such as strain gage, LVDT, RVDT, tachometer, etc. The noncontact type includes encoders, hall effect, capacitance, inductance, and interferometer type. They can also be classified based on the range of measurement. Usually the high-resolution type of sensors such as *hall effect, fiber optic inductance, capacitance*, and *strain gage* are suitable for only very small range (typically from 0.1 mm to 5 mm). The *differential transformers* on the other hand, have a much larger range with good resolution. *Interferometer* type sensors provide both very high resolution (in terms of microns) and large range of measurements (typically up to a meter). However, interferometer type sensors are bulky, expensive, and requires large set up time.

Among many linear displacement sensors, strain gage provides high resolution at low noise level and is least expensive. A typical resistance strain gage consists of resistive foil arranged as shown in the Fig. 16.2. A typical setup to measure the normal strain of a member loaded in tension is shown in Fig. 16.3. Strain gage 1 is bonded to the loading member whereas strain gage 2 is bonded to a second member made of same material, but not loaded. This arrangement compensates for any temperature effect. When the member is loaded, the gage 1 elongates thereby changing the resistance of the gage. The change in resistance is transformed into a change in voltage by the voltage-sensitive wheatstone bridge circuit. Assuming that the resistance of all four arms are equal initially, the change in output voltage (Δv_o) due to change in resistance (ΔR_1) of gage 1 is

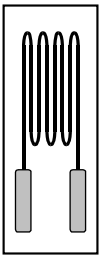



FIGURE 16.2 Bonded strain gage.

$$\frac{\Delta v_o}{v_i} = \frac{\Delta R_1/R}{4 + 2(\Delta R_1/R)}$$

Acceleration Sensors

Measurement of acceleration is important for systems subject to shock and vibration. Although acceleration can be derived from the time history data obtainable from linear or rotary sensors, the accelerometers whose output is directly proportional to the acceleration is preferred. Two common types include

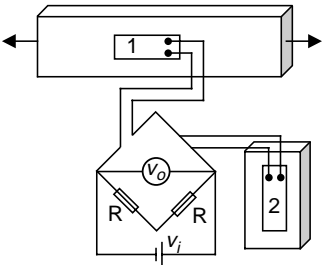



FIGURE 16.3 Experimental setup to measure normal strain using strain gages.

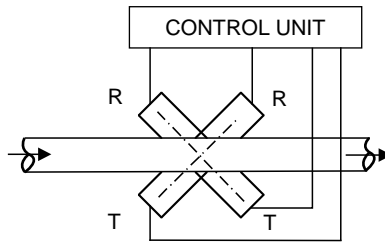



FIGURE 16.4 Ultrasonic flow sensor arrangement.

the *seismic mass* type and the *piezoelectric* accelerometer. The seismic mass type accelerometer is based on the relative motion between a mass and the supporting structure. The natural frequency of the seismic mass limits its use to low to medium frequency applications. The piezoelectric accelerometer, however, is compact and more suitable for high frequency applications.

Force, Torque, and Pressure Sensors

Among many type of force/torque sensors, the *strain gage dyanamometers* and *piezoelectric type* are most common. Both are available to measure force and/or torque either in one axis or multiple axes. The dynamometers make use of mechanical members that experiences elastic deflection when loaded. These types of sensors are limited by their natural frequency. On the other hand, the piezoelectric sensors are particularly suitable for dynamic loadings in a wide range of frequencies. They provide high stiffness, high resolution over a wide measurement range, and are compact.

Flow Sensors

Flow sensing is relatively a difficult task. The fluid medium can be liquid, gas, or a mixture of the two. Furthermore, the flow could be laminar or turbulent and can be a time-varying phenomenon. The *venturi meter* and *orifice plate* restrict the flow and use the pressure difference to determine the flow rate. The *pitot tube* pressure probe is another popular method of measuring flow rate. When positioned against the flow, they measure the total and static pressures. The flow velocity and in turn the flow rate can then be determined. The *rotameter* and the *turbine meters* when placed in the flow path, rotate at a speed proportional to the flow rate. The *electromagnetic flow meters* use noncontact method. Magnetic field is applied in the transverse direction of the flow and the fluid acts as the conductor to induce voltage proportional to the flow rate.

Ultrasonic flow meters measure fluid velocity by passing high-frequency sound waves through fluid. A schematic diagram of the ultrasonic flow meter is as shown in Fig. 16.4. The transmitters (T) provide the sound signal source. As the wave travels towards the receivers (R), its velocity is influenced by the velocity of the fluid flow due to the doppler effect. The control circuit compares the time to interpret the flow rate. This can be used for very high flow rates and can also be used for both upstream and downstream flow. The other advantage is that it can be used for corrosive fluids, fluids with abrasive particles, as it is like a noncontact sensor.

Temperature Sensors

A variety of devices are available to measure temperature, the most common of which are thermocouples, thermisters, resistance temperature detectors (RTD), and infrared types.

Thermocouples are the most versatile, inexpensive, and have a wide range (up to 1200∞C typical). A thermocouple simply consists of two dissimilar metal wires joined at the ends to create the sensing junction. When used in conjunction with a reference junction, the temperature difference between the reference junction and the actual temperature shows up as a voltage potential. *Thermisters* are semiconductor devices whose resistance changes as the temperature changes. They are good for very high sensitivity measurements in a limited range of up to 100∞C. The relationship between the temperature and the resistance is nonlinear. The *RTD*s use the phenomenon that the resistance of a metal changes with temperature. They are, however, linear over a wide range and most stable.

Infrared type sensors use the radiation heat to sense the temperature from a distance. These noncontact sensors can also be used to sense a field of vision to generate a thermal map of a surface.

Proximity Sensors

They are used to sense the proximity of an object relative to another object. They usually provide a on or off signal indicating the presence or absence of an object. *Inductance, capacitance, photoelectric*, and *hall effect* types are widely used as proximity sensors. Inductance proximity sensors consist of a coil wound around a soft iron core. The inductance of the sensor changes when a ferrous object is in its proximity. This change is converted to a voltage-triggered switch. Capacitance types are similar to inductance except the proximity of an object changes the gap and affects the capacitance. Photoelectric sensors are normally aligned with an infrared light source. The proximity of a moving object interrupts the light beam causing the voltage level to change. Hall effect voltage is produced when a current-carrying conductor is exposed to a transverse magnetic field. The voltage is proportional to transverse distance between the hall effect sensor and an object in its proximity.

Light Sensors

Light intensity and full field vision are two important measurements used in many control applications. *Phototransistors, photoresistors*, and *photodiodes* are some of the more common type of light intensity sensors. A common photoresistor is made of cadmium sulphide whose resistance is maximum when the sensor is in dark. When the photoresistor is exposed to light, its resistance drops in proportion to the intensity of light. When interfaced with a circuit as shown in Fig. 16.5 and balanced, the change in light intensity will show up as change in voltage. These sensors are simple, reliable, and cheap, used widely for measuring light intensity.

Smart Material Sensors

There are many new smart materials that are gaining more applications as sensors, especially in distributed sensing circumstances. Of these, *optic fibers, piezoelectric,* and *magnetostrictive* materials have found applications. Within these, optic fibers are most used.

Optic fibers can be used to sense strain, liquid level, force, and temperature with very high resolution. Since they are economical for use as *in situ* distributed sensors on large areas, they have found numerous applications in smart structure applications such as damage sensors, vibration sensors, and cure-monitoring sensors. These sensors use the inherent material (glass and silica) property of optical fiber to sense the environment. Figure 16.6 illustrates the basic principle of operation of an embedded optic fiber used to sense displacement, force, or temperature. The relative change in the transmitted intensity or spectrum is proportional to the change in the sensed parameter.

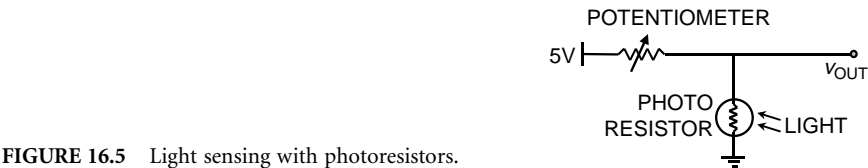



FIGURE 16.5 Light sensing with photoresistors.

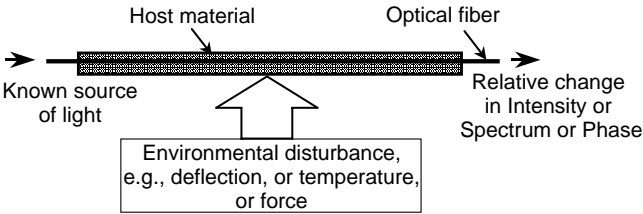



FIGURE 16.6 Principle of operation of optic fiber sensing.

Micro- and Nanosensors

Microsensors (sometimes also called MEMS) are the miniaturized version of the conventional macrosensors with improved performance and reduced cost. Silicon micromachining technology has helped the development of many microsensors and continues to be one of the most active research and development topics in this area.

Vision microsensors have found applications in medical technology. A *fiberscope* of approximately 0.2 mm in diameter has been developed to inspect flaws inside tubes. Another example is a *microtactile sensor*, which uses laser light to detect the contact between a catheter and the inner wall of blood vessels during insertion that has sensitivity in the range of 1 mN. Similarly, the progress made in the area of nanotechnology has fuelled the development of nanosensors. These are relatively new sensors that take one step further in the direction of miniaturization and are expected to open new avenues for sensing applications.

Selection Criteria

A number of static and dynamic factors must be considered in selecting a suitable sensor to measure the desired physical parameter. Following is a list of typical factors:

Range—Difference between the maximum and minimum value of the sensed parameter
Resolution—The smallest change the sensor can differentiate
Accuracy—Difference between the measured value and the true value
Precision—Ability to reproduce repeatedly with a given accuracy
Sensitivity—Ratio of change in output to a unit change of the input
Zero offset—A nonzero value output for no input
Linearity—Percentage of deviation from the best-fit linear calibration curve
Zero Drift—The departure of output from zero value over a period of time for no input
Response time—The time lag between the input and output
Bandwidth—Frequency at which the output magnitude drops by 3 dB
Resonance—The frequency at which the output magnitude peak occurs
Operating temperature—The range in which the sensor performs as specified
Deadband—The range of input for which there is no output
Signal-to-noise ratio—Ratio between the magnitudes of the signal and the noise at the output

Choosing a sensor that satisfies all the above to the desired specification is difficult, at best. For example, finding a position sensor with micrometer resolution over a range of a meter eliminates most of the sensors. Many times the lack of a cost-effective sensor necessitates redesigning the mechatronic system. It is, therefore, advisable to take a system level approach when selecting a sensor and avoid choosing it in isolation.

Once the above-referred functional factors are satisfied, a short list of sensors can be generated. The final selection will then depend upon the size, extent of signal conditioning, reliability, robustness, maintainability, and cost.

Signal Conditioning

Normally, the output from a sensor requires post processing of the signals before they can be fed to the controller. The sensor output may have to be demodulated, amplified, filtered, linearized, range quantized, and isolated so that the signal can be accepted by a typical analog-to-digital converter of the controller. Some sensors are available with integrated signal conditioners, such as the microsensors. All the electronics are integrated into one microcircuit and can be directly interfaced with the controllers.

Calibration

The sensor manufacturer usually provides the calibration curves. If the sensors are stable with no drift, there is no need to recalibrate. However, often the sensor may have to be recalibrated after integrating it with a signal conditioning system. This essentially requires that a known input signal is provided to

the sensor and its output recorded to establish a correct output scale. This process proves the ability to measure reliably and enhances the confidence.

If the sensor is used to measure a time-varying input, dynamic calibration becomes necessary. Use of sinusoidal inputs is the most simple and reliable way of dynamic calibration. However, if generating sinusoidal input becomes impractical (for example, temperature signals) then a step input can substitute for the sinusoidal signal. The transient behavior of step response should yield sufficient information about the dynamic response of the sensor.

16.2 Actuators

Actuators are basically the muscle behind a mechatronics system that accepts a control command (mostly in the form of an electrical signal) and produces a change in the physical system by generating force, motion, heat, flow, etc. Normally, the actuators are used in conjunction with the power supply and a coupling mechanism as shown in Fig. 16.7. The power unit provides either AC or DC power at the rated voltage and current. The coupling mechanism acts as the interface between the actuator and the physical system. Typical mechanisms include rack and pinion, gear drive, belt drive, lead screw and nut, piston, and linkages.

Classification

Actuators can be classified based on the type of energy as listed in Table 16.2. The table, although not exhaustive, lists all the basic types. They are essentially of electrical, electromechanical, electromagnetic, hydraulic, or pneumatic type. The new generations of actuators include smart material actuators, micro-actuators, and Nanoactuators.

Actuators can also be classified as *binary* and *continuous* based on the number of stable-state outputs. A relay with two stable states is a good example of a binary actuator. Similarly, a stepper motor is a good example of continuous actuator. When used for a position control, the stepper motor can provide stable outputs with very small incremental motion.

Principle of Operation

Electrical Actuators

Electrical switches are the choice of actuators for most of the on-off type control action. Switching devices such as *diodes, transistors, triacs, MOSFET,* and *relays* accept a low energy level command signal from the controller and switch on or off electrical devices such as motors, valves, and heating elements. For example, a MOSFET switch is shown in Fig. 16.8. The gate terminal receives the low energy control signal from the controller that makes or breaks the connection between the power supply and the actuator load. When switches are used, the designer must make sure that *switch bounce* problem is eliminated either by hardware or software.

Electromechanical Actuators

The most common electromechanical actuator is a motor that converts electrical energy to mechanical motion. Motors are the principal means of converting electrical energy into mechanical energy in industry. Broadly they can be classified as *DC motors, AC motors*, and *stepper motors*. DC motors operate on DC

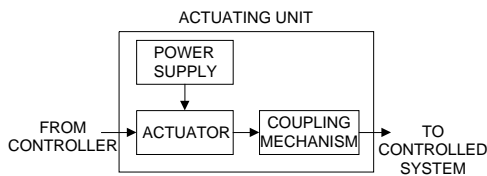



FIGURE 16.7 A typical actuating unit.

TABLE 16.2 Type of Actuators and Their Features

Actuator			Features
			Electrical
Diodes, thyristor, bipolar transistor, triacs, diacs, power MOSFET, solid state relay, etc.			Electronic type Very high frequency response Low power consumption
			Electromechanical
DC motor	Wound field	Separately excited	Speed can be controlled either by the voltage across the armature winding or by varying the field current
		Shunt	Constant-speed application
		Series	High starting torque, high acceleration torque, high speed with light load
		Compound	Low starting torque, good speed regulation Instability at heavy loads
	Permanent magnet	Conventional PM motor	High efficiency, high peak power, and fast response
		Moving-coil PM motor	Higher efficiency and lower inductance than conventional DC motor
		Torque motor	Designed to run for a long periods in a stalled or a low rpm condition
	Electronic commutation (brushless motor)		Fast response High efficiency, often exceeding 75% Long life, high reliability, no maintenance needed Low radio frequency interference and noise production
AC motor	AC induction motor		The most commonly used motor in industry Simple, rugged, and inexpensive
	AC synchronous motor		Rotor rotates at synchronous speed Very high efficiency over a wide range of speeds and loads Need an additional system to start
	Universal motor		Can operate in DC or AC Very high horsepower per pound ratio Relatively short operating life
Stepper motor	Hybrid		Change electrical pulses into mechanical movement Provide accurate positioning without feedback
	Variable reluctance		Low maintenance
			Electromagnetic
Solenoid type devices			Large force, short duration
Electromagnets, relay			On/off control
			Hydraulic and Pneumatic
Cylinder			Suitable for liner movement
Hydraulic motor	Gear type		Wide speed range
	Vane type		High horsepower output
	Piston type		High degree of reliability
Air motor	Rotary type		No electric shock hazard
	Reciprocating		Low maintenance
Valves	Directional control valves		
	Pressure control valves		
	Process control valves		
			Smart Material actuators
Piezoelectric & Electrostrictive			High frequency with small motion High voltage with low current excitation High resolution

(*continued*)

TABLE 16.2 Type of Actuators and Their Features (Continued)

Actuator	Features
Magnetostrictive	High frequency with small motion Low voltage with high current excitation
Shape Memory Alloy	Low voltage with high current excitation Low frequency with large motion
Electrorheological fluids	Very high voltage excitation Good resistance to mechanical shock and vibration Low frequency with large force
Micro- and Nanoactuators	
Micromotors	Suitable for micromechanical system
Microvalves	Can use available silicon processing technology, such as electrostatic motor
Micropumps	Can use any smart material

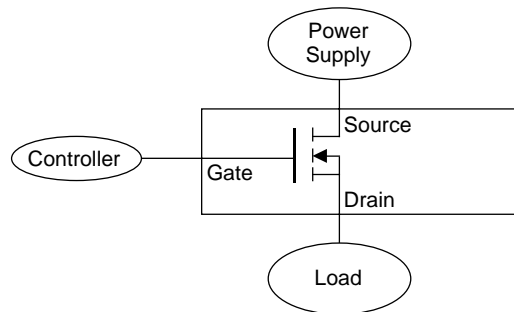



FIGURE 16.8 n-channel power MOSFET.

voltage and varying the voltage can easily control their speed. They are widely used in applications ranging from thousands of horsepower motors used in rolling mills to fractional horsepower motors used in automobiles (starter motors, fan motors, windshield wiper motors, etc.). Although they are costlier, they need DC power supply and require more maintenance compared to AC motors.

The governing equation of motion of a DC motor can be written as:

$$T = J\frac{d\omega}{dt} + T_L + T_{\text{loss}}$$

where T is torque, J is the total inertia, ω is the angular mechanical speed of the rotor, T_L is the torque applied to the motor shaft, and T_{loss} is the internal mechanical losses such as friction.

AC motors are the most popular since they use standard AC power, do not require brushes and commutator, and are therefore less expensive. AC motors can be further classified as the *induction motors, synchronous motors*, and *universal motors* according to their physical construction. The induction motor is simple, rugged, and maintenance free. They are available in many sizes and shapes based on number of phases used. For example, a three-phase induction motor is used in large-horsepower applications, such as pump drives, steel mill drives, hoist drives, and vehicle drives. The two-phase servomotor is used extensively in position control systems. Single-phase induction motors are widely used in many household appliances. The synchronous motor is one of the most efficient electrical motors in industry, so it is used in industry to reduce the cost of electrical power. In addition, synchronous motors rotate at synchronous speed, so they are also used in applications that require synchronous operations. The universal motors operate with either

AC or DC power supply. They are normally used in fractional horsepower application. The DC universal motor has the highest horsepower-per-pound ratio, but has a relatively short operating life.

The *stepper motor* is a discrete (incremental) positioning device that moves one step at a time for each pulse command input. Since they accept direct digital commands and produce a mechanical motion, the stepper motors are used widely in industrial control applications. They are mostly used in fractional horsepower applications. With the rapid progress in low cost and high frequency solid-state drives, they are finding increased applications.

Figure 16.9 shows a simplified unipolar stepper motor. The winding-1 is between the top and bottom stator pole, and the winding-2 is between the left and right motor poles. The rotor is a permanent magnet with six poles resulting in a single step angle of 30∞. With appropriate excitation of winding-1, the top stator pole becomes a north pole and the bottom stator pole becomes a south pole. This attracts the rotor into the position as shown. Now if the winding-1 is de-energized and winding-2 is energized, the rotor will turn 30∞. With appropriate choice of current flow through winding-2, the rotor can be rotated either clockwise or counterclockwise. By exciting the two windings in sequence, the motor can be made to rotate at a desired speed continuously.

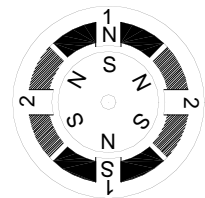



FIGURE 16.9 Unipolar stepper motor.

Electromagnetic Actuators

The *solenoid* is the most common electromagnetic actuator. A DC solenoid actuator consists of a soft iron core enclosed within a current carrying coil. When the coil is energized, a magnetic field is established that provides the force to push or pull the iron core. AC solenoid devices are also encountered, such as AC excitation relay.

A solenoid operated directional control valve is shown in Fig. 16.10. Normally, due to the spring force, the soft iron core is pushed to the extreme left position as shown. When the solenoid is excited, the soft iron core will move to the right extreme position thus providing the electromagnetic actuation.

Another important type is the *electromagnet*. The electromagnets are used extensively in applications that require large forces.

Hydraulic and Pneumatic Actuators

Hydraulic and pneumatic actuators are normally either *rotary motors* or *linear piston/cylinder* or *control valves*. They are ideally suited for generating very large forces coupled with large motion. Pneumatic actuators use air under pressure that is most suitable for low to medium force, short stroke, and high-speed applications. Hydraulic actuators use pressurized oil that is incompressible. They can produce very large forces coupled with large motion in a cost-effective manner. The disadvantage with the hydraulic actuators is that they are more complex and need more maintenance.

The rotary motors are usually used in applications where low speed and high torque are required. The cylinder/piston actuators are suited for application of linear motion such as aircraft flap control. Control valves in the form of directional control valves are used in conjunction with rotary motors and cylinders to control the fluid flow direction as shown in Fig. 16.10. In this solenoid operated directional control valve, the valve position dictates the direction motion of the cylinder/piston arrangement.

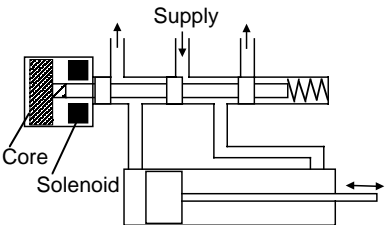



FIGURE 16.10 Solenoid operated directional control valve.

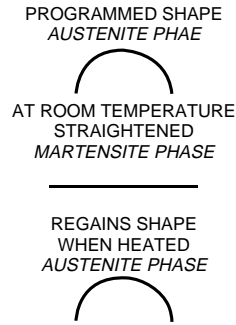



FIGURE 16.11 Phase changes of Shape Memory Alloy.

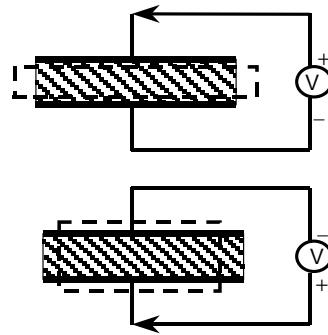



FIGURE 16.12 Piezoelectric actuator.

Smart Material Actuators

Unlike the conventional actuators, the smart material actuators typically become part of the load bearing structures. This is achieved by embedding the actuators in a distributed manner and integrating into the load bearing structure that could be used to suppress vibration, cancel the noise, and change shape. Of the many smart material actuators, *shape memory alloys, piezoelectric* (*PZT*), *magnetostrictive, Electrorheological fluids*, and *ion exchange polymers* are most common.

Shape Memory Alloys (SMA) are alloys of nickel and titanium that undergo phase transformation when subjected to a thermal field. The SMAs are also known as NITINOL for Nickel Titanium Naval Ordnance Laboratory. When cooled below a critical temperature, their crystal structure enters martensitic phase as shown in Fig. 16.11. In this state the alloy is plastic and can easily be manipulated. When the alloy is heated above the critical temperature (in the range of 50–80∞C), the phase changes to austenitic phase. Here the alloy resumes the shape that it formally had at the higher temperature. For example, a straight wire at room temperature can be made to regain its programmed semicircle shape when heated that has found applications in orthodontics and other tensioning devices. The wires are typically heated by passing a current (up to several amperes), 0 at very low voltage (2–10 V typical).

The PZT actuators are essentially piezocrystals with top and bottom conducting films as shown in Fig. 16.12. When an electric voltage is applied across the two conducting films, the crystal expands in the transverse direction as shown by the dotted lines. When the voltage polarity is reversed, the crystal contracts thereby providing bidirectional actuation. The interaction between the mechanical and electrical behavior of the piezoelectric materials can be expressed as:

$$T = c^E S - eE$$

where T is the stress, c^E is the elastic coefficients at constant electric field, S is the strain, e is the dielectric permitivity, and E is the electric field.

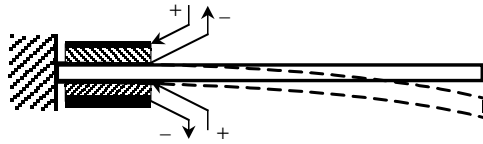



FIGURE 16.13 Vibration of beam using piezoelectric actuators.

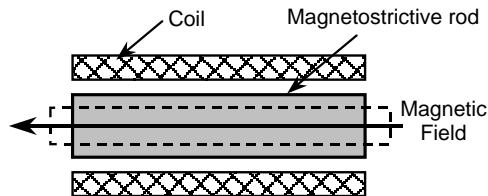



FIGURE 16.14 Magnetostrictive rod actuator.

One application of these actuators is as shown in Fig. 16.13. The two piezoelectric patches are excited with opposite polarity to create transverse vibration in the cantilever beam. These actuators provide high bandwidth (0–10 kHz typical) with small displacement. Since there are no moving parts to the actuator, it is compact and ideally suited for micro and nano actuation. Unlike the bidirectional actuation of piezoelectric actuators, the *electrostriction* effect is a second-order effect, i.e., it responds to an electric field with unidirectional expansion regardless of polarity.

Magnetostrictive material is an alloy of terbium, dysprosium, and iron that generates mechanical strains up to 2000 microstrain in response to applied magnetic fields. They are available in the form of rods, plates, washers, and powder. Figure 16.14 shows a typical magnetostrictive rod actuator that is surrounded by a magnetic coil. When the coil is excited, the rod elongates in proportion to the intensity of the magnetic field established. The magnetomechanical relationship is given as:

$$\varepsilon = S^H \sigma + dH$$

where, ε is the strain, S^H the compliance at constant magnetic filed, σ the stress, d the magnetostriction constant, and H the magnetic field intensity.

Ion exchange polymers exploit the electro-osmosis phenomenon of the natural ionic polymers for purposes of actuation. When a voltage potential is applied across the cross-linked polyelectrolytic network, the ionizable groups attain a net charge generating a mechanical deformation. These types of actuators have been used to develop artificial muscles and artificial limbs. The primary advantage is their capacity to produce large deformation with a relatively low voltage excitation.

Micro- and Nanoactuators

Microactuators, also called micromachines, microelectromechanical system (MEMS), and microsystems are the tiny mobile devices being developed utilizing the standard microelectronics processes with the integration of semiconductors and machined micromechanical elements. Another definition states that any device produced by assembling extremely small functional parts of around 1–15 mm is called a micromachine.

In *electrostatic motors*, electrostatic force is dominant, unlike the conventional motors that are based on magnetic forces. For smaller micromechanical systems the electrostatic forces are well suited as an actuating force. Figure 16.15 shows one type of electrostatic motor. The rotor is an annular disk with uniform permitivity and conductivity. In operation, a voltage is applied to the two conducting parallel

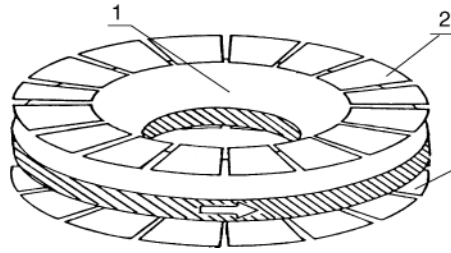



FIGURE 16.15 Electrostatic motor: 1-rotor, 2-stator electrodes.

plates separated by an insulation layer. The rotor rotates with a constant velocity between the two coplanar concentric arrays of stator electrodes.

Selection Criteria

The selection of the proper actuator is more complicated than selection of the sensors, primarily due to their effect on the dynamic behavior of the overall system. Furthermore, the selection of the actuator dominates the power needs and the coupling mechanisms of the entire system. The coupling mechanism can sometimes be completely avoided if the actuator provides the output that can be directly interfaced to the physical system. For example, choosing a linear motor in place of a rotary motor can eliminate the need of a coupling mechanism to convert rotary motion to linear motion.

In general, the following performance parameters must be addressed before choosing an actuator for a specific need:

Continuous power output—The maximum force/torque attainable continuously without exceeding the temperature limits
Range of motion—The range of linear/rotary motion
Resolution—The minimum increment of force/torque attainable
Accuracy—Linearity of the relationship between the input and output
Peak force/torque—The force/torque at which the actuator stalls
Heat dissipation—Maximum wattage of heat dissipation in continuous operation
Speed characteristics—Force/torque versus speed relationship
No load speed—Typical operating speed/velocity with no external load
Frequency response—The range of frequency over which the output follows the input faithfully, applicable to linear actuators
Power requirement—Type of power (AC or DC), number of phases, voltage level, and current capacity

In addition to the above-referred criteria, many other factors become important depending upon the type of power and the coupling mechanism required. For example, if a rack- and-pinion coupling mechanism is chosen, the backlash and friction will affect the resolution of the actuating unit.

17

Fundamentals of Time and Frequency

Michael A. Lombardi
National Institute of Standards and Technology

17.1 Introduction

Time and frequency standards supply three basic types of information: *time-of-day*, *time interval*, and *frequency*. Time-of-day information is provided in hours, minutes, and seconds, but often also includes the *date* (month, day, and year). A device that displays or records time-of-day information is called a *clock*. If a clock is used to label when an event happened, this label is sometimes called a *time tag* or *time stamp*. Date and time-of-day can also be used to ensure that events are *synchronized*, or happen at the same time.

Time interval is the duration or elapsed time between two events. The standard unit of time interval is the second(s). However, many engineering applications require the measurement of shorter time intervals, such as milliseconds (1 ms = 10^{-3} s), microseconds (1 μs = 10^{-6} s), nanoseconds (1 ns = 10^{-9} s), and picoseconds (1 ps = 10^{-12} s). Time is one of the seven base physical quantities, and the second is one of seven base units defined in the International System of Units (SI). The definitions of many other physical quantities rely upon the definition of the second. The second was once defined based on the earth's rotational rate or as a fraction of the tropical year. That changed in 1967 when the era of atomic time keeping formally began. The current definition of the SI second is:

> The duration of 9,192,631,770 periods of the radiation corresponding to the transition between two hyperfine levels of the ground state of the cesium-133 atom.

Frequency is the rate of a repetitive event. If T is the period of a repetitive event, then the frequency f is its reciprocal, $1/T$. Conversely, the period is the reciprocal of the frequency, $T = 1/f$. Since the period is a time interval expressed in seconds (s), it is easy to see the close relationship between time interval and frequency. The standard unit for frequency is the hertz (Hz), defined as events or cycles per second. The frequency of electrical signals is often measured in multiples of hertz, including kilohertz (kHz), megahertz (MHz), or gigahertz (GHz), where 1 kHz equals one thousand (10^3) events per second, 1 MHz

TABLE 17.1 Uncertainties of Physical Realizations of the Base SI Units

SI Base Unit	Physical Quantity	Uncertainty
Candela	Luminous intensity	1×10^{-4}
Kelvin	Temperature	3×10^{-7}
Mole	Amount of substance	8×10^{-8}
Ampere	Electric current	4×10^{-8}
Kilogram	Mass	1×10^{-8}
Meter	Length	1×10^{-12}
Second	Time interval	1×10^{-15}

equals one million (10^6) events per second, and 1 GHz equals one billion (10^9) events per second. A device that produces frequency is called an *oscillator*. The process of setting multiple oscillators to the same frequency is called *syntonization*.

Of course, the three types of time and frequency information are closely related. As mentioned, the standard unit of time interval is the second. By counting seconds, we can determine the date and the time-of-day. And by counting events or cycles per second, we can measure frequency.

Time interval and frequency can now be measured with less uncertainty and more resolution than any other physical quantity. Today, the best time and frequency standards can realize the SI second with uncertainties of $\cong 1 \times 10^{-15}$. Physical realizations of the other base SI units have much larger uncertainties, as shown in Table 17.1 [1–5].

Coordinated Universal Time (UTC)

The world's major metrology laboratories routinely measure their time and frequency standards and send the measurement data to the Bureau International des Poids et Measures (BIPM) in Sevres, France. The BIPM averages data collected from more than 200 atomic time and frequency standards located at more than 40 laboratories, including the National Institute of Standards and Technology (NIST). As a result of this averaging, the BIPM generates two time scales, International Atomic Time (TAI), and Coordinated Universal Time (UTC). These time scales realize the SI second as closely as possible.

UTC runs at the same frequency as TAI. However, it differs from TAI by an integral number of seconds. This difference increases when *leap seconds* occur. When necessary, leap seconds are added to UTC on either June 30 or December 31. The purpose of adding leap seconds is to keep atomic time (UTC) within ±0.9 s of an older time scale called UT1, which is based on the rotational rate of the earth. Leap seconds have been added to UTC at a rate of slightly less than once per year, beginning in 1972 [3,5].

Keep in mind that the BIPM maintains TAI and UTC as "paper" time scales. The major metrology laboratories use the published data from the BIPM to steer their clocks and oscillators and generate real-time versions of UTC. Many of these laboratories distribute their versions of UTC via radio signals, which are discussed in section 17.4.

You can think of UTC as the ultimate standard for time-of-day, time interval, and frequency. Clocks synchronized to UTC display the same hour, minute, and second all over the world (and remain within one second of UT1). Oscillators syntonized to UTC generate signals that serve as reference standards for time interval and frequency.

17.2 Time and Frequency Measurement

Time and frequency measurements follow the conventions used in other areas of metrology. The frequency standard or clock being measured is called the *device under test* (*DUT*). A measurement compares the DUT to a *standard* or *reference*. The standard should outperform the DUT by a specified ratio, called the *test uncertainty ratio* (*TUR*). Ideally, the TUR should be 10:1 or higher. The higher the ratio, the less averaging is required to get valid measurement results.

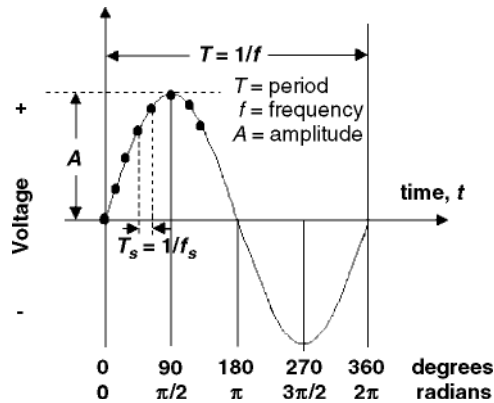



FIGURE 17.1 An oscillating sine wave.

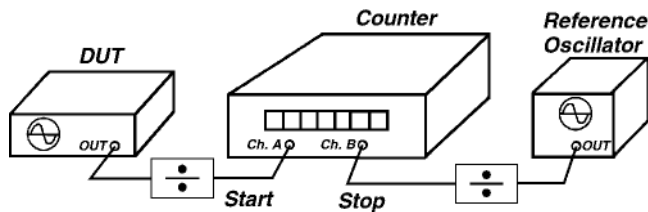



FIGURE 17.2 Measurement using a time interval counter.

The test signal for time measurements is usually a pulse that occurs once per second (1 pps). The pulse width and polarity varies from device to device, but TTL levels are commonly used. The test signal for frequency measurements is usually at a frequency of 1 MHz or higher, with 5 or 10 MHz being common. Frequency signals are usually sine waves, but can also be pulses or square waves. If the frequency signal is an oscillating sine wave, it might look like the one shown in Fig. 17.1. This signal produces one cycle (360∞ or 2π radians of phase) in one period. The signal amplitude is expressed in volts, and must be compatible with the measuring instrument. If the amplitude is too small, it might not be able to drive the measuring instrument. If the amplitude is too large, the signal must be attenuated to prevent overdriving the measuring instrument.

This section examines the two main specifications of time and frequency measurements—*accuracy* and *stability*. It also discusses some instruments used to measure time and frequency.

Accuracy

Accuracy is the degree of conformity of a measured or calculated value to its definition. Accuracy is related to the offset from an ideal value. For example, *time offset* is the difference between a measured on-time pulse and an ideal on-time pulse that coincides exactly with UTC. *Frequency offset* is the difference between a measured frequency and an ideal frequency with zero uncertainty. This ideal frequency is called the *nominal frequency*.

Time offset is usually measured with a *time interval counter* (*TIC*), as shown in Fig. 17.2. A TIC has inputs for two signals. One signal starts the counter and the other signal stops it. The time interval between the start and stop signals is measured by counting cycles from the time base oscillator. The resolution of a low cost TIC is limited to the period of its time base. For example, a TIC with a 10-MHz time base oscillator would have a resolution of 100 ns. More elaborate TICs use interpolation schemes to detect parts of a time base cycle and have much higher resolution—1 ns resolution is commonplace, and 20 ps resolution is available.

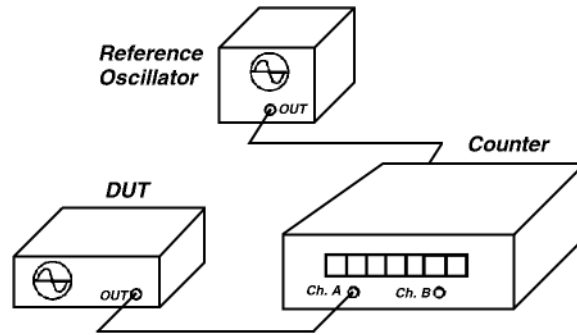



FIGURE 17.3 Measurement using a frequency counter.

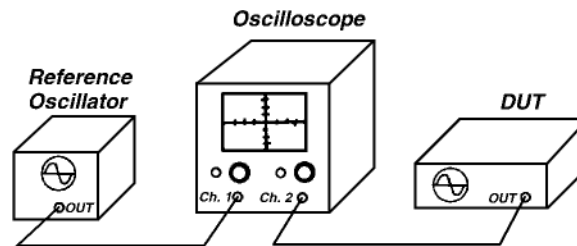



FIGURE 17.4 Phase comparison using an oscilloscope.

Frequency offset can be measured in either the *frequency domain* or *time domain*. A simple frequency domain measurement involves directly counting and displaying the frequency output of the DUT with a *frequency counter*. The reference for this measurement is either the counter's internal time base oscillator, or an external time base (Fig. 17.3). The counter's resolution, or the number of digits it can display, limits its ability to measure frequency offset. For example, a 9-digit frequency counter can detect a frequency offset no smaller than 0.1 Hz at 10 MHz (1×10^{-8}). The frequency offset is determined as

$$f(\text{offset}) = \frac{f_{\text{measured}} - f_{\text{nominal}}}{f_{\text{nominal}}}$$

where f_{measured} is the reading from the frequency counter, and f_{nominal} is the frequency labeled on the oscillator's nameplate, or specified output frequency.

Frequency offset measurements in the time domain involve a *phase comparison* between the DUT and the reference. A simple phase comparison can be made with an oscilloscope (Fig. 17.4). The oscilloscope will display two sine waves (Fig. 17.5). The top sine wave represents a signal from the DUT, and the bottom sine wave represents a signal from the reference. If the two frequencies were exactly the same, their phase relationship would not change and both would appear to be stationary on the oscilloscope display. Since the two frequencies are not exactly the same, the reference appears to be stationary and the DUT signal moves. By measuring the rate of motion of the DUT signal we can determine its frequency offset. Vertical lines have been drawn through the points where each sine wave passes through zero. The bottom of the figure shows bars whose width represents the phase difference between the signals. In this case the phase difference is increasing, indicating that the DUT is lower in frequency than the reference.

Measuring high accuracy signals with an oscilloscope is impractical, since the phase relationship between signals changes very slowly and the resolution of the oscilloscope display is limited. More precise phase comparisons can be made with a TIC, using a setup similar to Fig. 17.2. If the two input signals have the same frequency, the time interval will not change. If the two signals have different frequencies,

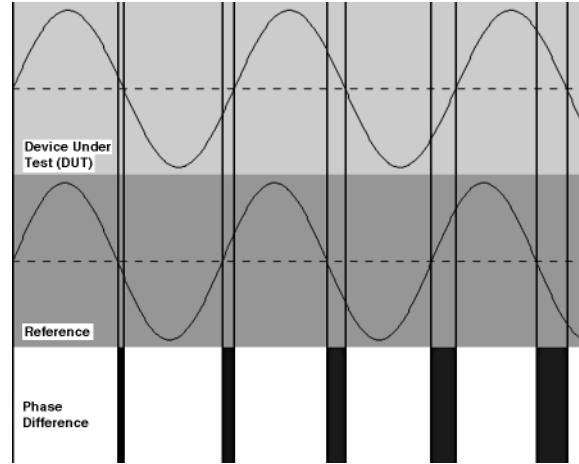



FIGURE 17.5 Two sine waves with a changing phase relationship.

the time interval will change, and the rate of change is the frequency offset. The resolution of a TIC determines the smallest frequency change that it can detect without averaging. For example, a low cost TIC with a single-shot resolution of 100 ns can detect frequency changes of 1×10^{-7} in 1 s. The current limit for TIC resolution is about 20 ps, which means that a frequency change of 2×10^{-11} can be detected in 1 s. Averaging over longer intervals can improve the resolution to <1 ps in some units [6].

Since standard frequencies like 5 or 10 MHz are not practical to measure with a TIC, *frequency dividers* (shown in Fig. 17.2) or *frequency mixers* are used to convert the test frequency to a lower frequency. Divider systems are simpler and more versatile, since they can be easily built or programmed to accommodate different frequencies. Mixer systems are more expensive, require more hardware including an additional reference oscillator, and can often measure only one input frequency (e.g., 10 MHz), but they have a higher signal-to-noise ratio than divider systems.

If dividers are used, measurements are made from the TIC, but instead of using these measurements directly, we determine the rate of change from reading to reading. This rate of change is called the *phase deviation*. We can estimate frequency offset as follows:

$$f(\text{offset}) = \frac{-\Delta t}{T}$$

where Δt is the amount of phase deviation, and T is the measurement period.

To illustrate, consider a measurement of +1 μs of phase deviation over a measurement period of 24 h. The unit used for measurement period (h) must be converted to the unit used for phase deviation (μs). The equation becomes

$$f(\text{offset}) = \frac{-\Delta t}{T} = \frac{-1\ \mu\text{s}}{86{,}400{,}000{,}000\ \mu\text{s}} = -1.16 \times 10^{-11}$$

As shown, a device that accumulates 1 μs of phase deviation/day has a frequency offset of -1.16×10^{-11} with respect to the reference. This simple example requires only two time interval readings to be made, and Δt is simply the difference between the two readings. Often, multiple readings are taken and the frequency offset is estimated by using least squares linear regression on the data set, and obtaining Δt from the slope of the least squares line. This information is usually presented as a phase plot, as shown in Fig. 17.6. The device under test is high in frequency by exactly 1×10^{-9}, as indicated by a phase deviation of 1 ns/s [2,7,8].

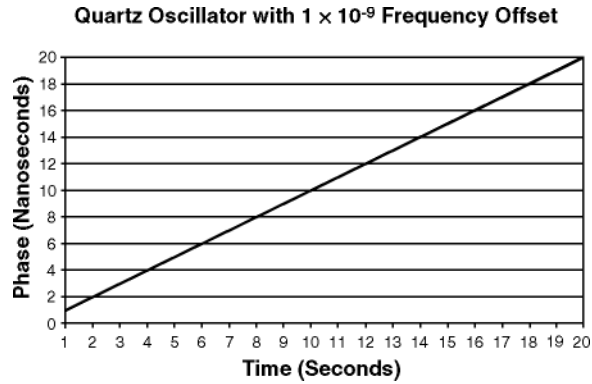



FIGURE 17.6 A sample phase plot.

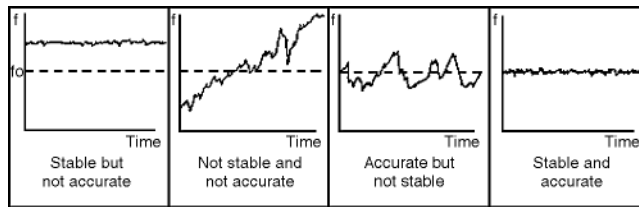



FIGURE 17.7 The relationship between accuracy and stability.

Dimensionless frequency offset values can be converted to units of frequency (Hz) if the nominal frequency is known. To illustrate this, consider an oscillator with a nominal frequency of 5 MHz and a frequency offset of $+1.16 \times 10^{-11}$. To find the frequency offset in hertz, multiply the nominal frequency by the offset:

$$(5 \times 10^{6})\ (+1.16 \times 10^{-11}) = 5.80 \times 10^{-5} = +0.0000580 \text{ Hz}$$

Then, add the offset to the nominal frequency to get the actual frequency:

$$5{,}000{,}000 \text{ Hz} + 0.0000580 \text{ Hz} = 5{,}000{,}000.0000580 \text{ Hz}$$

Stability

Stability indicates how well an oscillator can produce the same time or frequency offset over a given time interval. It doesn't indicate whether the time or frequency is "right" or "wrong," but only whether it *stays the same*. In contrast, accuracy indicates how well an oscillator has been set on time or on frequency. To understand this difference, consider that a stable oscillator that needs adjustment might produce a frequency with a large offset. Or, an unstable oscillator that was just adjusted might temporarily produce a frequency near its nominal value. Figure 17.7 shows the relationship between accuracy and stability.

Stability is defined as the statistical estimate of the frequency or time fluctuations of a signal over a given time interval. These fluctuations are measured with respect to a mean frequency or time offset. *Short-term* stability usually refers to fluctuations over intervals less than 100 s. *Long-term* stability can refer to measurement intervals greater than 100 s, but usually refers to periods longer than 1 day.

Stability estimates can be made in either the frequency domain or time domain, and can be calculated from a set of either frequency offset or time interval measurements. In some fields of measurement, stability is estimated by taking the standard deviation of the data set. However, standard deviation only

works with stationary data, where the results are time independent, and the noise is *white*, meaning that it is evenly distributed across the frequency band of the measurement. Oscillator data is usually nonstationary, since it contains time dependent noise contributed by the frequency offset. With stationary data, the mean and standard deviation will converge to particular values as more measurements are made. With nonstationary data, the mean and standard deviation never converge to any particular values. Instead, there is a moving mean that changes each time we add a measurement.

For these reasons, a non-classical statistic is often used to estimate stability in the time domain. This statistic is sometimes called the *Allan variance*, but since it is the square root of the variance, its proper name is the *Allan deviation*. The equation for the Allan deviation ($\sigma_y(\tau)$) is

$$\sigma_y(\tau) = \sqrt{\frac{1}{2(M-1)}\sum_{i=1}^{M-1}(y_{i+1}-y_i)^2}$$

where y_i is a set of frequency offset measurements containing y_1, y_2, y_3, and so on, M is the number of values in the y_i series, and the data are equally spaced in segments τ seconds long. Or

$$\sigma_y(\tau) = \sqrt{\frac{1}{2(N-2)\tau^2}\sum_{i=1}^{N-2}[x_{i+2}-2x_{i+1}+x_i]^2}$$

where x_i is a set of phase measurements in time units containing x_1, x_2, x_3, and so on, N is the number of values in the x_i series, and the data are equally spaced in segments τ seconds long. Note that while standard deviation subtracts the mean from each measurement before squaring their summation, the Allan deviation subtracts the previous data point. This differencing of successive data points removes the time dependent noise contributed by the frequency offset.

An Allan deviation graph is shown in Fig. 17.8. It shows the stability of the device improving as the averaging period (τ) gets longer, since some noise types can be removed by averaging. At some point, however, more averaging no longer improves the results. This point is called the *noise floor*, or the point where the remaining noise consists of nonstationary processes such as flicker noise or random walk. The device measured in Fig. 17.8 has a noise floor of $\sim 5 \times 10^{-11}$ at $\tau = 100$ s.

Practically speaking, a frequency stability graph also tells us how long we need to average to get rid of the noise contributed by the reference and the measurement system. The noise floor provides some indication of the amount of averaging required to obtain a TUR high enough to show us the true frequency

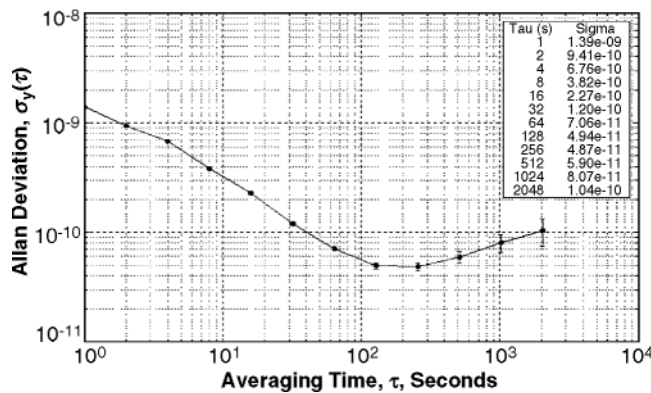



FIGURE 17.8 A frequency stability graph.

TABLE 17.2 Statistics Used to Estimate Time and Frequency Stability and Noise Types

Name	Mathematical Notation	Description
Allan deviation	$\sigma_y(\tau)$	Estimates frequency stability. Particularly suited for intermediate- to long-term measurements.
Modified Allan deviation	MOD $\sigma_y(\tau)$	Estimates frequency stability. Unlike the normal Allan deviation, it can distinguish between white and flicker phase noise, which makes it more suitable for short-term stability estimates.
Time deviation	$\sigma_x(\tau)$	Used to measure time stability. Clearly identifies both white and flicker phase noise, the noise types of most interest when measuring time or phase.
Total deviation	$\sigma_{y,\,\mathrm{TOTAL}}(\tau)$	Estimates frequency stability. Particularly suited for long-term estimates where τ exceeds 10% of the total data sample.

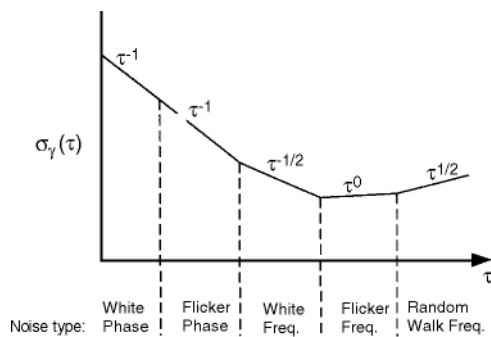



FIGURE 17.9 Using a frequency stability graph to identify noise types.

offset of the DUT. If the DUT is an atomic oscillator (section 17.4) and the reference is a radio controlled transfer standard (section 17.5) we might have to average for 24 h or longer to have confidence in the measurement result.

Five noise types are commonly discussed in the time and frequency literature: *white phase, flicker phase, white frequency, flicker frequency*, and *random walk frequency*. The slope of the Allan deviation line can help identify the amount of averaging needed to remove these noise types (Fig. 17.9). The first type of noise to be removed by averaging is phase noise, or the rapid, random fluctuations in the phase of the signal. Ideally, only the device under test would contribute phase noise to the measurement, but in practice, some phase noise from the measurement system and reference needs to be removed through averaging. Note that the Allan deviation does not distinguish between white phase noise and flicker phase noise. Table 17.2 shows several other statistics used to estimate stability and identify noise types for various applications.

Identifying and eliminating sources of oscillator noise can be a complex subject, but plotting the first order differences of a set of time domain measurements can provide a basic understanding of how noise is removed by averaging. Figure 17.10 was made using a segment of the data from the stability graph in Fig. 17.8. It shows phase plots dominated by white phase noise (1 s averaging), white frequency noise (64 s averages), flicker frequency noise (256 s averages), and random walk frequency (1024 s averages). Note that the white phase noise plot has a 2 ns scale, and the other plots use a 100 ps scale [8–12].

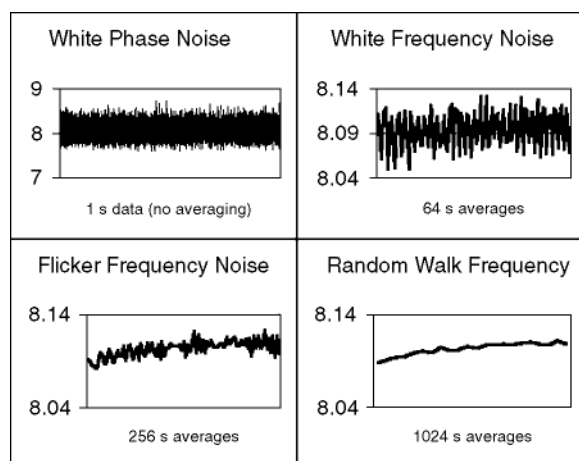



FIGURE 17.10 Phase plots of four noise types.

17.3 Time and Frequency Standards

All time and frequency standards are based on a *periodic event* that repeats at a constant rate. The device that produces this event is called a *resonator*. In the simple case of a pendulum clock, the pendulum is the resonator. Of course, a resonator needs an energy source before it can move back and forth. Taken together, the energy source and resonator form an *oscillator*. The oscillator runs at a rate called the *resonance frequency*. For example, a clock's pendulum can be set to swing back and forth at a rate of once per second. Counting one complete swing of the pendulum produces a time interval of 1 s. Counting the total number of swings creates a *time scale* that establishes longer time intervals, such as minutes, hours, and days. The device that does the counting and displays or records the results is called a *clock*. Table 17.3 shows how the frequency uncertainty of a clock's resonator corresponds to the timing uncertainty of a clock.

Throughout history, clock designers have searched for more stable resonators, and the evolution of time and frequency standards is summarized in Table 17.4. The uncertainties listed for modern standards represent current (year 2001) devices, and not the original prototypes. Note that the performance of time and frequency standards has improved by 13 orders of magnitude in the past 700 years, and by about nine orders of magnitude in the past 100 years.

The stability of time and frequency standards is closely related to their quality factor, or Q. The Q of an oscillator is its resonance frequency divided by its resonance width. The resonance frequency is the natural frequency of the oscillator. The resonance width is the range of possible frequencies where the oscillator will oscillate. A high-Q resonator will not oscillate at all unless it is near its resonance frequency. Obviously, a high resonance frequency and a narrow resonance width are both advantages when seeking a high Q. Generally speaking, the higher the Q, the more stable the oscillator, since a high Q means that an oscillator will stay close to its natural resonance frequency.

This section begins by discussing quartz oscillators, which achieve the highest Q of any mechanical-type device. It then discusses oscillators with much higher Q factors, based on the atomic resonance of rubidium and cesium. Atomic oscillators use the quantized energy levels in atoms and molecules as the source of their resonance. The laws of quantum mechanics dictate that the energies of a bound system, such as an atom, have certain discrete values. An electromagnetic field at a particular frequency can boost an atom from one energy level to a higher one. Or, an atom at a high energy level can drop to a lower level by emitting energy. The resonance frequency (f) of an atomic oscillator is the difference between

Table 17.3 Relationship of Frequency Uncertainty to Time Uncertainty

Frequency Uncertainty	Measurement Period	Time Uncertainty
$\pm 1.00 \times 10^{-3}$	1 s	± 1 ms
$\pm 1.00 \times 10^{-6}$	1 s	± 1 μs
$\pm 1.00 \times 10^{-9}$	1 s	± 1 ns
$\pm 2.78 \times 10^{-7}$	1 h	± 1 ms
$\pm 2.78 \times 10^{-10}$	1 h	± 1 μs
$\pm 2.78 \times 10^{-13}$	1 h	± 1 ns
$\pm 1.16 \times 10^{-8}$	1 day	± 1 ms
$\pm 1.16 \times 10^{-11}$	1 day	± 1 μs
$\pm 1.16 \times 10^{-14}$	1 day	± 1 ns

TABLE 17.4 The Evolution of Time and Frequency Standards

Standard	Resonator	Date of Origin	Timing Uncertainty (24 h)	Frequency Uncertainty (24 h)
Sundial	Apparent motion of the sun	3500 B.C.	NA	NA
Verge escapement	Verge and foliet mechanism	14th century	15 min	1×10^{-2}
Pendulum	Pendulum	1656	10 s	1×10^{-4}
Harrison chronometer (H4)	Spring and balance wheel	1759	350 ms	4×10^{-6}
Shortt pendulum	Two pendulums, slave and master	1921	10 ms	1×10^{-7}
Quartz crystal	Quartz crystal	1927	10 μs	1×10^{-10}
Rubidium gas cell	^{87}Rb resonance (6,834,682,608 Hz)	1958	100 ns	1×10^{-12}
Cesium beam	^{133}Cs resonance (9,192,631,770 Hz)	1952	1 ns	1×10^{-14}
Hydrogen maser	Hydrogen resonance (1,420,405,752 Hz)	1960	1 ns	1×10^{-14}
Cesium fountain	^{133}Cs resonance (9,192,631,770 Hz)	1991	100 ps	1×10^{-15}

the two energy levels divided by Planck's constant (h):

$$f = \frac{E_2 - E_1}{h}$$

The principle underlying the atomic oscillator is that since all atoms of a specific element are identical, they should produce exactly the same frequency when they absorb or release energy. In theory, the atom is a perfect "pendulum" whose oscillations are counted to measure time interval. The discussion of atomic oscillators is limited to devices that are commercially available, and excludes the primary and experimental standards found in laboratories such as NIST. Table 17.5 provides a summary [1,4,8].

Quartz Oscillators

Quartz crystal oscillators are by far the most common time and frequency standards. An estimated two billion (2×10^9) quartz oscillators are manufactured annually. Most are small devices built for wrist-watches, clocks, and electronic circuits. However, they are also found inside test and measurement equipment, such as counters, signal generators, and oscilloscopes; and interestingly enough, inside every atomic oscillator.

TABLE 17.5 Summary of Oscillator Types

Oscillator Type	Quartz (TCXO)	Quartz (OCXO)	Rubidium	Commercial Cesium Beam	Hydrogen Maser
Q	10^4 to 10^6	3.2×10^6 (5 MHz)	10^7	10^8	10^9
Resonance frequency	Various	Various	6.834682608 GHz	9.192631770 GHz	1.420405752 GHz
Leading cause of failure	None	None	Rubidium lamp (life expectancy >15 years)	Cesium beam tube (life expectancy of 3 to 25 years)	Hydrogen depletion (life expectancy >7 years)
Stability, $\sigma_y(\tau)$, $\tau = 1$ s	1×10^{-8} to 1×10^{-9}	1×10^{-12}	5×10^{-11} to 5×10^{-12}	5×10^{-11} to 5×10^{-12}	1×10^{-12}
Noise floor, $\sigma_y(\tau)$	1×10^{-9} ($\tau = 1$ to 10^2 s)	1×10^{-12} ($\tau = 1$ to 10^2 s)	1×10^{-12} ($\tau = 10^3$ to 10^5 s)	1×10^{-14} ($\tau = 10^5$ to 10^7 s)	1×10^{-15} ($\tau = 10^3$ to 10^5 s)
Aging/year	5×10^{-7}	5×10^{-9}	1×10^{-10}	None	~ 1×10^{-13}
Frequency offset after warm-up	1×10^{-6}	1×10^{-8} to 1×10^{-10}	5×10^{-10} to 5×10^{-12}	5×10^{-12} to 1×10^{-14}	1×10^{-12} to 1×10^{-13}
Warm-Up period	<10 s to 1×10^{-6}	<5 min to 1×10^{-8}	<5 min to 5×10^{-10}	30 min to 5×10^{-12}	24 h to 1×10^{-12}

A quartz crystal inside the oscillator is the resonator. It can be made of either natural or synthetic quartz, but all modern devices use synthetic quartz. The crystal strains (expands or contracts) when a voltage is applied. When the voltage is reversed, the strain is reversed. This is known as the *piezoelectric effect*. Oscillation is sustained by taking a voltage signal from the resonator, amplifying it, and feeding it back to the resonator. The rate of expansion and contraction is the resonance frequency and is determined by the cut and size of the crystal. The output frequency of a quartz oscillator is either the fundamental resonance or a multiple of the resonance, called an *overtone frequency*. Most high stability units use either the third or fifth overtone to achieve a high Q. Overtones higher than fifth are rarely used because they make it harder to tune the device to the desired frequency. A typical Q for a quartz oscillator ranges from 10^4 to 10^6. The maximum Q for a high stability quartz oscillator can be estimated as $Q = 1.6 \times 10^7/f$, where f is the resonance frequency in megahertz.

Environmental changes due to temperature, humidity, pressure, and vibration can change the resonance frequency of a quartz crystal, but there are several designs that reduce these environmental effects. The *oven-controlled crystal oscillator* (*OCXO*) encloses the crystal in a temperature-controlled chamber called an oven. When an OCXO is turned on, it goes through a "warm-up" period while the temperatures of the crystal resonator and its oven stabilize. During this time, the performance of the oscillator continuously changes until it reaches its normal operating temperature. The temperature within the oven then remains constant, even when the outside temperature varies. An alternate solution to the temperature problem is the *temperature-compensated crystal oscillator* (*TCXO*). In a TCXO, the signal from a temperature sensor is used to generate a correction voltage that is applied to a voltage-variable reactance, or varactor. The varactor then produces a frequency change equal and opposite to the frequency change produced by temperature. This technique does not work as well as oven control, but is less expensive. Therefore, TCXOs are used when high stability over a wide temperature range is not required.

Quartz oscillators have excellent short-term stability. An OCXO might be stable ($\sigma_y(\tau)$, at $\tau = 1$ s) to 1×10^{-12}. The limitations in short-term stability are due mainly to noise from electronic components in the oscillator circuits. Long-term stability is limited by *aging*, or a change in frequency with time due to internal changes in the oscillator. Aging is usually a nearly linear change in the resonance frequency that can be either positive or negative, and occasionally, a reversal in direction of aging occurs. Aging has many possible causes including a build-up of foreign material on the crystal, changes in the oscillator circuitry,

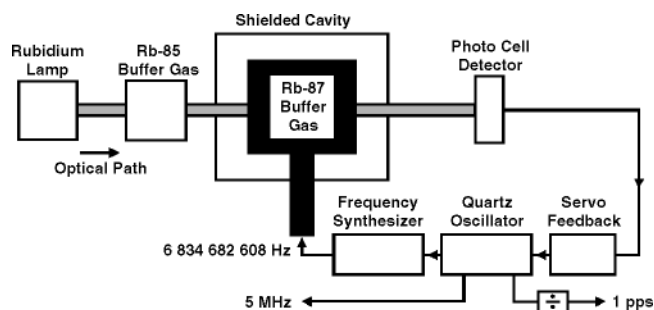



FIGURE 17.11 Rubidium oscillator.

or changes in the quartz material or crystal structure. A high quality OCXO might age at a rate of <5 × 10^{-9} per year, while a TCXO might age 100 times faster.

Due to aging and environmental factors such as temperature and vibration, it is hard to keep even the best quartz oscillators within 1×10^{-10} of their nominal frequency without constant adjustment. For this reason, atomic oscillators are used for applications that require better long-term accuracy and stability [4,13,14].

Rubidium Oscillators

Rubidium oscillators are the lowest priced members of the atomic oscillator family. They operate at 6,834,682,608 Hz, the resonance frequency of the rubidium atom (^{87}Rb), and use the rubidium frequency to control the frequency of a quartz oscillator. A microwave signal derived from the crystal oscillator is applied to the ^{87}Rb vapor within a cell, forcing the atoms into a particular energy state. An optical beam is then pumped into the cell and is absorbed by the atoms as it forces them into a separate energy state. A photo cell detector measures how much of the beam is absorbed, and its output is used to tune a quartz oscillator to a frequency that maximizes the amount of light absorption. The quartz oscillator is then locked to the resonance frequency of rubidium, and standard frequencies are derived from the quartz oscillator and provided as outputs (Fig. 17.11).

Rubidium oscillators continue to get smaller and less expensive, and offer perhaps the best price-to-performance ratio of any oscillator. Their long-term stability is much better than that of a quartz oscillator and they are also smaller, more reliable, and less expensive than cesium oscillators.

The Q of a rubidium oscillator is about 10^7. The shifts in the resonance frequency are due mainly to collisions of the rubidium atoms with other gas molecules. These shifts limit the long-term stability. Stability ($\sigma_y(\tau)$, at $\tau = 1$ s) is typically 1×10^{-11}, and about 1×10^{-12} at 1 day. The frequency offset of a rubidium oscillator ranges from 5×10^{-10} to 5×10^{-12} after a warm-up period of a few minutes or hours, so they meet the accuracy requirements of most applications without adjustment.

Cesium Oscillators

Cesium oscillators are *primary frequency standards* since the SI second is defined from the resonance frequency of the cesium atom (^{133}Cs), which is 9,192,631,770 Hz. A properly working cesium oscillator should be close to its nominal frequency without adjustment, and there should be no change in frequency due to aging.

Commercially available oscillators use *cesium beam* technology. Inside a cesium oscillator, ^{133}Cs atoms are heated to a gas in an oven. Atoms from the gas leave the oven in a high-velocity beam that travels through a vacuum tube toward a pair of magnets. The magnets serve as a gate that allows only atoms of a particular magnetic energy state to pass into a microwave cavity, where they are exposed to a microwave

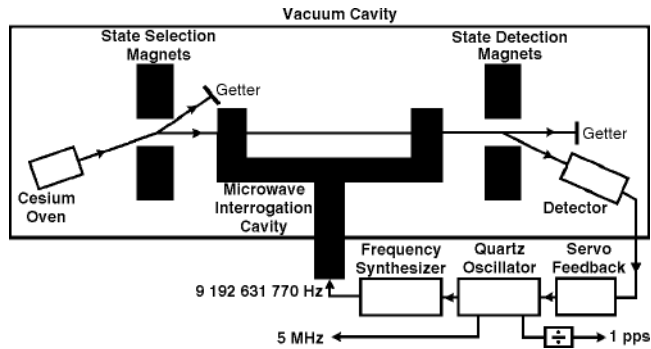



FIGURE 17.12 Cesium beam oscillator.

frequency derived from a quartz oscillator. If the microwave frequency matches the resonance frequency of cesium, the cesium atoms change their magnetic energy state.

The atomic beam then passes through another magnetic gate near the end of the tube. Those atoms that changed their energy state while passing through the microwave cavity are allowed to proceed to a detector at the end of the tube. Atoms that did not change state are deflected away from the detector. The detector produces a feedback signal that continually tunes the quartz oscillator in a way that maximizes the number of state changes so that the greatest number of atoms reaches the detector. Standard output frequencies are derived from the locked quartz oscillator (Fig. 17.12).

The Q of a commercial cesium standard is a few parts in 10^8. The beam tube is typically <0.5 m in length, and the atoms travel at velocities of >100 m/s inside the tube. This limits the observation time to a few milliseconds, and the resonance width to a few hundred hertz. Stability ($\sigma_y(\tau)$, at $\tau = 1$ s) is typically 5×10^{-12} and reaches a noise floor near 1×10^{-14} at about 1 day, extending out to weeks or months. The frequency offset is typically near 1×10^{-12} after a warm-up period of 30 min.

17.4 Time and Frequency Transfer

Many applications require clocks or oscillators at different locations to be set to the same time (*synchronization*), or the same frequency (*syntonization*). *Time and frequency transfer* techniques are used to compare and adjust clocks and oscillators at different locations. Time and frequency transfer can be as simple as setting your wristwatch to an audio time signal, or as complex as controlling the frequency of oscillators in a network to parts in 10^{13}.

Time and frequency transfer can use signals broadcast through many different media, including coaxial cables, optical fiber, radio signals (at numerous places in the radio spectrum), telephone lines, and the Internet. Synchronization requires both an on-time pulse and a time code. Syntonization requires extracting a stable frequency from the broadcast. The frequency can come from the carrier itself, or from a time code or other information modulated onto the carrier.

This section discusses both the fundamentals of time and frequency transfer and the radio signals used as calibration references. Table 17.6 provides a summary.

Fundamentals of Time and Frequency Transfer

Signals used for time and frequency transfer are generally referenced to atomic oscillators that are steered to agree as closely as possible with UTC. Information is sent from a transmitter (A) to a receiver (B) and is delayed by τ_{ab}, commonly called the *path delay* (Fig. 17.13).

To illustrate path delay, consider a radio signal broadcast over a path 1000 km long. Since radio signals travel at the speed of light (~3.3 μs/km), we can calibrate the path by applying a 3.3-ms correction to

TABLE 17.6 Summary of Time and Frequency Transfer Signals and Methods

Signal or Link	Receiving Equipment	Time Uncertainty (24 h)	Frequency Uncertainty (24 h)
Dial-Up Computer Time Service	Computer, client software, modem, and phone line	<15 ms	Not recommended for frequency measurements
Internet Time Service	Computer, client software, and Internet connection	<1 s	Not recommended for frequency measurements
HF Radio (3 to 30 MHz)	HF receiver and antenna	1 to 20 ms	10^{-6} to 10^{-9}
LF Radio (30 to 300 kHz)	LF receiver and antenna	1 to 100 μs	10^{-10} to 10^{-12}
Global Positioning System (GPS)	GPS receiver antenna	<20 ns	$<2 \times 10^{-13}$

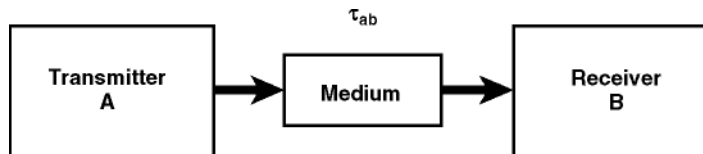



FIGURE 17.13 One-way time and frequency transfer.

our measurement. Of course, for many applications the path delay is simply ignored. For example, if our goal is simply to synchronize a computer clock within 1 s of UTC, there is no need to worry about a 100-ms path delay through a network. And, of course, path delay is not important to frequency transfer systems, since on-time pulses are not required. Instead, frequency transfer requires only a stable path where the delays remain relatively constant.

More sophisticated transfer systems estimate and remove all or part of the path delay. This is usually done in one of two ways. The first way is to estimate τ_{ab} and send the time out early by this amount. For example, if τ_{ab} is at least 20 ms for all users, the time can be sent 20 ms early. This advancement of the timing signal removes at least some of the delay for all users.

A better technique is to compute τ_{ab} and to apply a correction to the received signal. A correction for τ_{ab} can be computed if the position of both the transmitter and receiver are known. If the transmitter is stationary, a constant can be used for the transmitter position. If the transmitter is moving (a satellite, for example) it must broadcast its position in addition to broadcasting time. The Global Positioning System (GPS) provides the best of both worlds—each GPS satellite broadcasts its position and the receiver can use coordinates from multiple satellites to compute its own position.

The transmitted information often includes a *time code* so that a clock can be set to the correct time-of-day. Most time codes contain the UTC hour, minute, and second; the month, day, and year; and advance warning of daylight saving time and leap seconds.

Radio Time and Frequency Transfer Signals

There are many types of radio receivers designed to receive time and frequency signals. Some are designed primarily to produce time-of-day information or an on-time pulse, others are designed to output standard frequencies, and some can be used for both time and frequency transfer. The following sections look at three types of time and frequency radio signals that distribute UTC—high frequency (HF), low frequency (LF), and GPS satellite signals.

HF Radio Signals (Including WWV and WWVH)

High frequency (HF) radio broadcasts occupy the radio spectrum from 3 to 30 MHz. These signals are commonly used for time and frequency transfer at moderate performance levels. Some HF broadcasts provide audio time announcements and digital time codes. Other broadcasts simply provide a carrier frequency for use as a reference.

HF time and frequency stations include NIST radio stations WWV and WWVH. WWV is located near Fort Collins, Colorado, and WWVH is on the island of Kauai, Hawaii. Both stations broadcast continuous time and frequency signals on 2.5, 5, 10, and 15 MHz, and WWV also broadcasts on 20 MHz. All frequencies broadcast the same program, and at least one frequency should be usable at all times. The stations can also be heard by telephone; dial (303) 499-7111 for WWV or (808) 335-4363 for WWVH.

WWV and WWVH signals can be used in one of three modes:

- The audio portion of the broadcast includes seconds pulses or ticks, standard audio frequencies, and voice announcements of the UTC hour and minute. WWV uses a male voice, and WWVH uses a female voice.
- A binary time code is sent on a 100 Hz subcarrier at a rate of 1 bit per second. The time code contains the hour, minute, second, year, day of year, leap second and Daylight Saving Time (DST) indicators, and a UT1 correction. This code can be read and displayed by radio clocks.
- The carrier frequency can be used as a reference for the calibration of oscillators. This is done most often with the 5 and 10 MHz carrier signals, since they match the output frequencies of standard oscillators.

The time broadcast by WWV and WWVH will be late when it arrives at the user's location. The time offset depends upon the receiver's distance from the transmitter, but should be <15 ms in the continental United States. A good estimate of the time offset requires knowledge of HF radio propagation. Most users receive a signal that has traveled up to the ionosphere and was then reflected back to earth. Since the height of the ionosphere changes throughout the day, the path delay also changes. Path delay variations limit the received frequency uncertainty to parts in 10^9 when averaged for 1 day.

HF radio stations such as WWV and WWVH are useful for low level applications, such as the manual synchronization of analog and digital clocks, simple frequency calibrations, and calibrations of stop watches and timers. However, LF and GPS signals are better choices for more demanding applications [2,7,15].

LF Radio Signals (Including WWVB)

Before the advent of satellites, low frequency (LF) signals were the method of choice for time and frequency transfer. While the use of LF signals has diminished in the laboratory, they still have two major advantages—they can often be received indoors without an external antenna and several stations broadcast a time code. This makes them ideal for many consumer electronic products that display time-of-day information.

Many time and frequency stations operate in the LF band from 30 to 300 kHz (Table 17.7). The performance of the received signal is influenced by the path length and signal strength. Path length is important because the signal is divided into ground wave and sky wave. The ground wave signal is more stable. Since it travels the shortest path between the transmitter and receiver, it arrives first and its path delay is much easier to estimate. The sky wave is reflected from the ionosphere and produces results similar to those obtained with HF reception. Short paths make it possible to continuously track the ground wave. Longer paths produce a mixture of sky wave and ground wave. And over very long paths, only sky wave reception is possible.

Signal strength is also important. If the signal is weak, the receiver might search for a new cycle of the carrier to track. Each time the receiver adjusts its tracking point by one cycle, it introduces a phase step equal to the period of the carrier. For example, a cycle slip on a 60 kHz carrier introduces a 16.67 μs phase step. However, a strong ground wave signal can produce very good results. An LF receiver that

TABLE 17.7 LF Time and Frequency Broadcast Stations

Call Sign	Country	Frequency (kHz)	Always On?
DCF77	Germany	77.5	Yes
DGI	Germany	177	Yes
HBG	Switzerland	75	Yes
JG2AS	Japan	40	Yes
MSF	United Kingdom	60	Yes
RBU	Russia	66.666	No
RTZ	Russia	50	Yes
TDF	France	162	Yes
WWVB	United States	60	Yes

continuously tracks the same cycle of a ground wave signal can transfer frequency with an uncertainty of about 1×10^{-12} when averaged for 1 day.

NIST operates LF radio station WWVB from Fort Collins, Colorado at a transmission frequency of 60 kHz. The station broadcasts 24 h per day, with an effective radiated output power of 50 kW. The WWVB time code is synchronized with the 60 kHz carrier and contains the year, day of year, hour, minute, second, and flags that indicate the status of daylight saving time, leap years, and leap seconds. The time code is received and displayed by wristwatches, alarm clocks, wall clocks, and other consumer electronic products [2,7,15].

Global Positioning System (GPS)

The GPS is a navigation system developed and operated by the U.S. Department of Defense (DoD) that is usable nearly anywhere on the earth. The system consists of a constellation of at least 24 satellites that orbit the earth at a height of 20,200 km in six fixed planes inclined 55∞ from the equator. The orbital period is 11 h 58 m, which means that each satellite will pass over the same place on earth twice per day. By processing signals received from the satellites, a GPS receiver can determine its position with an uncertainty of <10 m.

The satellites broadcast on two carrier frequencies, L1 at 1575.42 MHz and L2 at 1227.6 MHz. Each satellite broadcasts a spread spectrum waveform, called a *pseudo random noise* (*PRN*) code on L1 and L2, and each satellite is identified by the PRN code it transmits. There are two types of PRN codes. The first type is a *coarse acquisition* (*C/A*) code with a chipping rate of 1023 chips per millisecond. The second is a *precision* (*P*) code with a chipping rate of 10230 chips per millisecond. The C/A code is broadcast on L1, and the P code is broadcast on both L1 and L2. GPS reception is line-of-sight, which means that the receiving antenna must have a clear view of the sky [16].

Each satellite carries either rubidium or cesium oscillators, or a combination of both. These oscillators are steered from DoD ground stations and are referenced to the United States Naval Observatory time scale, UTC (USNO), which by agreement is always maintained within 100 ns of UTC (NIST). The oscillators provide the reference for both the carrier and the code broadcasts.

GPS signals now dominate the world of high performance time and frequency transfer, since they provide reliable reception and exceptional results with minimal effort. A GPS receiver can automatically compute its latitude, longitude, and altitude from position data received from the satellites. The receiver can then calibrate the radio path and synchronize its on-time pulse. In addition to the on-time pulse, many receivers provide standard frequencies such as 5 or 10 MHz by steering an OCXO or rubidium oscillator using the satellite signals. GPS receivers also produce time-of-day and date information.

A GPS receiver calibrated for equipment delays has a timing uncertainty of <20 ns relative to UTC (NIST), and the frequency uncertainty is often $<2 \times 10^{-13}$ when averaged for 1 day. Figure 17.14 shows an Allan deviation plot of the output of a low cost GPS receiver. The stability is near 1×10^{-13} after about 1 day of averaging.

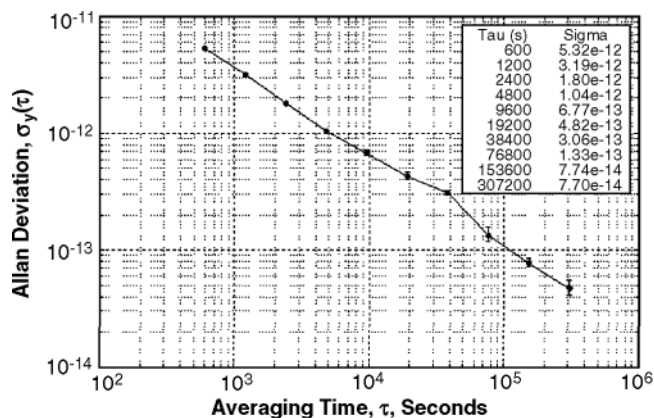



FIGURE 17.14 Frequency stability of GPS receiver.

17.5 Closing

As noted earlier, time and frequency standards and measurements have improved by about nine orders of magnitude in the past 100 years. This rapid advance has made many new products and technologies possible. While it is impossible to predict what the future holds, we can be certain that oscillator Qs will continue to increase, measurement uncertainties will continue to decrease, and new technologies will continue to emerge.

References

1. Jespersen, J., and Fitz-Randolph, J., *From Sundials to Atomic Clocks: Understanding Time and Frequency,* 2nd ed., Dover, Mineola, New York, 1999.
2. Kamas, G., and Lombardi, M. A., *Time and Frequency Users Manual,* NIST Special Publication 559, U.S. Government Printing Office, Washington, DC, 1990.
3. Levine, J., Introduction to time and frequency metrology, *Rev. Sci. Instrum.*, 70, 2567, 1999.
4. Hackman, C., and Sullivan, D. B., Eds., *Time and Frequency Measurement,* American Association of Physics Teachers, College Park, Maryland, 1996.
5. ITU Radiocommunication Study Group 7, *Selection and Use of Precise Frequency and Time Systems,* International Telecommunications Union, Geneva, Switzerland, 1997.
6. Novick, A. N., Lombardi, M. A., Zhang, V. S., and Carpentier, A., A high performance multi-channel time interval counter with an integrated GPS receiver, in *Proc. 31st Annu. Precise Time and Time Interval (PTTI) Meeting,* Dana Point, California, p. 561, 1999.
7. Lombardi, M. A., Time measurement and frequency measurement, in *The Measurement, Instrumentation, and Sensors Handbook,* Webster, J. G., Eds., CRC Press, Boca Raton, Florida, 1999, chap. 18–19.
8. Sullivan, D. B., Allan, D. W., Howe, D. A., and Walls, F. L., Eds., *Characterization of Clocks and Oscillators,* NIST Technical Note 1337, U.S. Government Printing Office, Washington, DC, 1990.
9. Jespersen, J., Introduction to the time domain characterization of frequency standards, in *Proc. 23rd Annu. Precise Time and Time Interval (PTTI) Meeting,* Pasadena, California, p. 83, 1991.
10. IEEE Standards Coordinating Committee 27, *IEEE Standard Definitions of Physical Quantities for Fundamental Frequency and Time Metrology—Random Instabilities,* Institute of Electrical and Electronics Engineers, New York, 1999.
11. Walls, F. L., and Ferre-Pikal, E. S., Measurement of frequency, phase noise, and amplitude noise, in *Wiley Encyclopedia of Electrical and Electronics Engineering,* John Wiley and Sons, New York, 1999, 12, 459.

12. Howe, D. A., An extension of the Allan variance with increased confidence at long term, *IEEE Int. Freq. Control Symp.,* 321, 1995.
13. Vig, J. R., Introduction to quartz frequency standards, *Army Research and Development Technical Report,* SLCET-TR-92-1, October 1992.
14. Hewlett-Packard Company, *Fundamentals of Quartz Oscillators,* HP Application Note 200-2, 1997.
15. Carr, J. J., *Elements of Electronic Instrumentation and Measurement,* 3rd ed., Prentice-Hall, NJ, 1996.
16. Hoffmann-Wellenhof, B., Lichtenegger, H., and Collins, J., *GPS: Theory and Practice, 3rd* ed., Springer-Verlag, New York, 1994.

18

Sensor and Actuator Characteristics

Joey Parker
University of Alabama

Mechatronic systems use a variety of sensors and actuators to measure and manipulate mechanical, electrical, and thermal systems. Sensors have many characteristics that affect their measurement capabilities and their suitability for each application. Analog sensors have an output that is continuous over a finite region of inputs. Examples of analog sensors include potentiometers, LVDTs (linear variable differential transformers), load cells, and thermistors. Digital sensors have a fixed or countable number of different output values. A common digital sensor often found in mechatronic systems is the incremental encoder. An analog sensor output conditioned by an analog-to-digital converter (ADC) has the same digital output characteristics, as seen in Fig. 18.1.

18.1 Range

The range (or span) of a sensor is the difference between the minimum (or most negative) and maximum inputs that will give a valid output. Range is typically specified by the manufacturer of the sensor. For example, a common type K thermocouple has a range of 800∞C (from -50∞C to 750∞C). A ten-turn potentiometer would have a range of 3600 degrees.

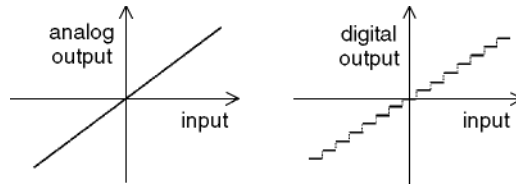



FIGURE 18.1 Analog and digital sensor outputs.

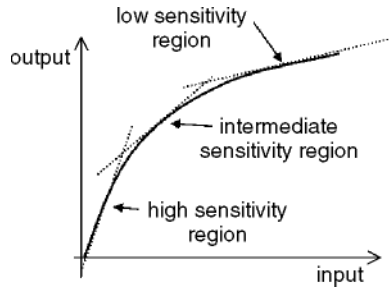



FIGURE 18.2 Sensor sensitivity.

18.2 Resolution

The resolution of a sensor is the smallest increment of input that can be reliably detected. Resolution is also frequently known as the least count of the sensor. Resolution of digital sensors is easily determined. A 1024 ppr (pulse per revolution) incremental encoder would have a resolution of

$$\frac{1\ \text{revolution}}{1024\ \text{pulses}} \times \frac{360\ \text{degrees}}{1\ \text{revolution}} = 0.3516\frac{\text{degrees}}{\text{pulse}}$$

The resolution of analog sensors is usually limited only by low-level electrical noise and is often much better than equivalent digital sensors.

18.3 Sensitivity

Sensor sensitivity is defined as the change in output per change in input. The sensitivity of digital sensors is closely related to the resolution. The sensitivity of an analog sensor is the slope of the output versus input line. A sensor exhibiting truly linear behavior has a constant sensitivity over the entire input range. Other sensors exhibit nonlinear behavior where the sensitivity either increases or decreases as the input is changed, as shown in Fig. 18.2.

18.4 Error

Error is the difference between a measured value and the true input value. Two classifications of errors are bias (or systematic) errors and precision (or random) errors. Bias errors are present in all measurements made with a given sensor, and cannot be detected or removed by statistical means. These bias errors can be further subdivided into

- calibration errors (a zero or null point error is a common type of bias error created by a nonzero output value when the input is zero),
- loading errors (adding the sensor to the measured system changes the system), and
- errors due to sensor sensitivity to variables other than the desired one (e.g., temperature effects on strain gages).

18.5 Repeatability

Repeatability (or reproducibility) refers to a sensor's ability to give identical outputs for the same input. Precision (or random) errors cause a lack of repeatability. Fortunately, precision errors can be accounted for by averaging several measurements or other operations such as low-pass filtering. Electrical noise and hysteresis (described later) both contribute to a loss of repeatability.

18.6 Linearity and Accuracy

The accuracy of a sensor is inversely proportional to error, i.e., a highly accurate sensor produces low errors. Many manufacturers specify accuracy in terms of the sensor's linearity. A least-squares straight-line fit between all output measurements and their corresponding inputs determines the nominal output of the sensor. Linearity (or accuracy) is specified as a percentage of full scale (maximum valid input), as shown in Fig. 18.3, or as a percentage of the sensor reading, as shown in Fig. 18.4. Figures 18.3 and 18.4 show both of these specifications for 10% linearity, which is much larger than most actual sensors.

Accuracy and precision are two terms that are frequently confused. Figure 18.5 shows four sets of histograms for ten measurements of angular velocity of an actuator turning at a constant 100 rad/s. The first set of data shows a high degree of precision (low standard deviation) and repeatability, but the

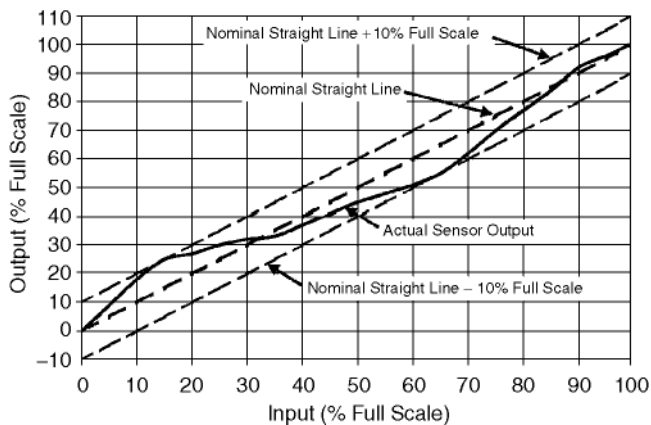



FIGURE 18.3 Linearity specified at full scale.

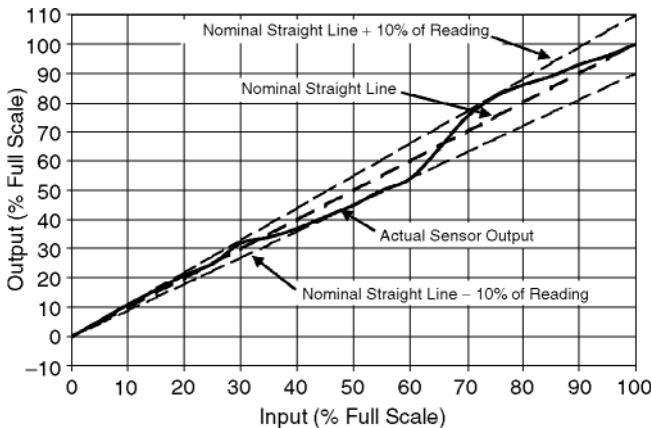



FIGURE 18.4 Linearity specified at reading.

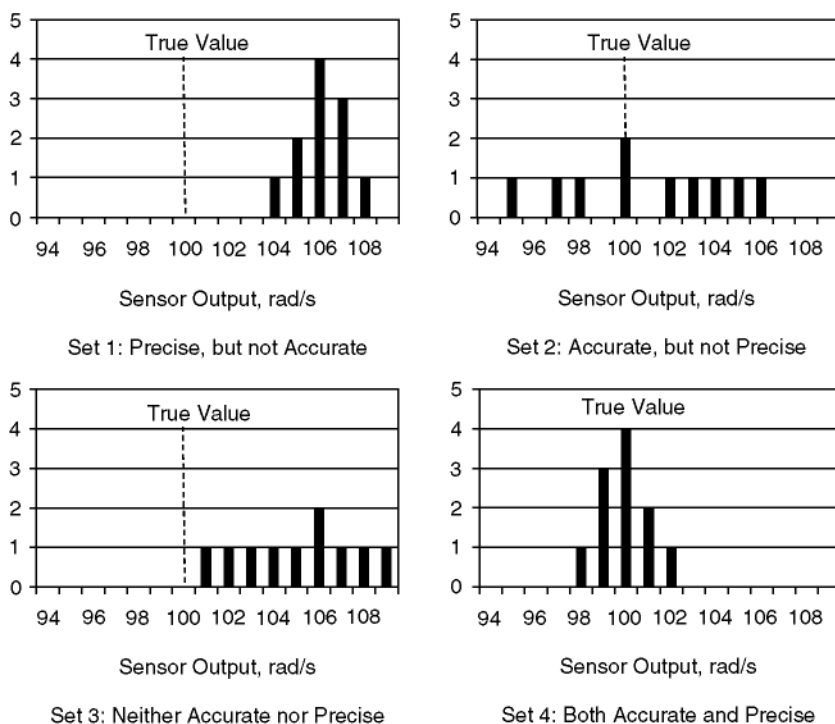



FIGURE 18.5 Examples of accuracy and precision.

average accuracy is poor. The second set of data shows a low degree of precision (high standard deviation), but the average accuracy is good. The third set of data shows both low precision and low accuracy, while the fourth set of data shows both high precision, high repeatability, and high accuracy.

18.7 Impedance

Impedance is the ratio of voltage and current flow for a sensor. For a simple resistive sensor (such as a strain gage or a thermistor), the impedance Z is the same as the resistance R, which has units of ohms (W),

$$Z_R = \frac{V}{I} = R$$

For more complicated sensors, impedance includes the effects of capacitance, C, and inductance, L. Inclusion of these terms makes the impedance frequency sensitive, but the units remain ohms:

$$Z_C = \frac{V}{I} = \frac{1}{jC\omega} \quad \text{and} \quad Z_L = \frac{V}{I} = jL\omega$$

where $j = \sqrt{-1}$ is the imaginary number and ω is the driving frequency. The impedance form is particularly nice for analyzing simple circuits, as parallel and series inductances can be treated just like resistances. Two types of impedance are important in sensor applications: input impedance and output impedance. Input impedance is a measure of how much current must be drawn to power a sensor (or signal conditioning circuit). Input impedance is frequently modeled as a resistor in parallel with the

input terminals. High input impedance is desirable, since the device will then draw less current from the source. Oscilloscopes and data acquisition equipment frequently have input impedances of 1 MΩ or more to minimize this current draw. Output impedance is a measure of a sensor's (or signal conditioning circuit's) ability to provide current for the next stage of the system. Output impedance is frequently modeled as a resistor in series with the sensor output. Low output impedance is desirable, but is often not available directly from a sensor. Piezoelectric sensors in particular have high output impedances and cannot source much current (typically micro-amps or less). Op-amp circuits are frequently used to buffer sensor outputs for this reason. Op-amp circuits (especially voltage followers) provide nearly ideal circumstances for many sensors, since they have high input impedance but can substantially lower output impedance.

18.8 Nonlinearities

Linear systems have the property of superposition. If the response of the system to input A is output A, and the response to input B is output B, then the response to input C (= input A + input B) will be output C (= output A + output B). Many real systems will exhibit linear or nearly linear behavior over some range of operation. Therefore, linear system analysis is correct, at least over these portions of a system's operating envelope. Unfortunately, most real systems have nonlinearities that cause them to operate outside of this linear region, and many common assumptions about system behavior, such as superposition, no longer apply. Several nonlinearities commonly found in mechatronic systems include static and coulomb friction, eccentricity, backlash (or hysteresis), saturation, and deadband.

18.9 Static and Coulomb Friction

In classic linear system analysis, friction forces are assumed to be proportional to velocity, i.e., viscous friction. With an actuator velocity of zero, there should be no friction. In reality, a small amount of static (no velocity) or Coulomb friction is almost always present, even in roller or ball type anti-friction bearings. A typical plot of friction force vs. velocity is given in Fig. 18.6. Note that the static friction force can assume any value between some upper and lower limit at zero velocity. Static friction has two primary effects on mechatronic systems:

1. Some of the actuator torque or force is wasted overcoming friction forces, which leads to inefficiency from an energy viewpoint.
2. As the actuator moves the system to its final location, the velocity approaches zero and the actuator force/torque will approach a value that exactly balances frictional and gravity loads. Since static friction can assume any value at zero velocity, the actuator will come to slightly different final resting positions each time—depending on the final value of static friction. This effect contributes to some loss of repeatability in mechatronic systems.

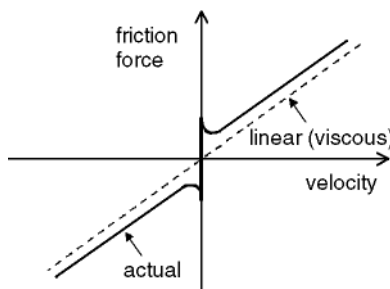



FIGURE 18.6 Static and Coulomb friction.

18.10 Eccentricity

The ideal relationships for gears, pulleys, and chain drives assume that the point of gear contact remains at a fixed distance from the center of rotation for each gear. In reality, the true center of the gears pitch circle and the center of rotation will be separated by a small amount, known as the eccentricity. Small tooth-to-tooth errors can also cause local variations in the pitch circle radius. The combination of these two effects can lead to a nonlinear geometrical relationship between two gears like that of Fig. 18.7, where the nonlinear behavior is greatly exaggerated for clarity. Eccentricity impacts the accuracy of position measurements made on the input side of the gear pair, as the output gear is not exactly where the sensor measurement indicates.

18.11 Backlash

If two otherwise perfect gears are not mounted on a center-to-center distance that exactly matches the sum of the pitch radii, there will be a small clearance, or backlash, between the teeth. When the input gear reverses direction, a small rotation is required before this clearance is removed and the output gear begins to move. Gear backlash is just one of many phenomena that can be characterized as hysteresis, as shown in Fig. 18.8. Clearance between shafts and bearings can cause hysteretic effects also. Backlash exhibits effects similar to those for eccentricity, i.e., a loss of repeatability, particularly when approaching a measured point from different directions. The gear backlash problem is so prevalent and potentially harmful that many manufacturers go to great lengths to minimize or reduce the effect:

- gears mounted closer together than the theoretically ideal spacing,
- split "anti-backlash" gears that are spring loaded to force teeth to maintain engagement at all times,
- external spring-loaded mounts for one of the gears to force engagement, or
- specially designed gears with anti-backlash features.

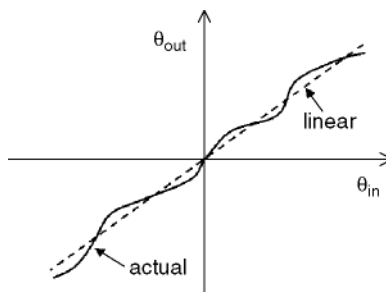



FIGURE 18.7 Gear eccentricity.

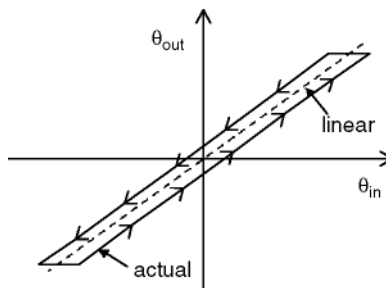



FIGURE 18.8 Gear backlash.

18.12 Saturation

All real actuators have some maximum output capability, regardless of the input. This violates the linearity assumption, since at some point the input command can be increased without significantly changing the output; see Fig. 18.9. This type of nonlinearity must be considered in mechatronic control system design, since maximum velocity and force or torque limitations affect system performance. Control systems modeled with linear system theory must be carefully tested or analyzed to determine the impact of saturation on system performance.

18.13 Deadband

Another nonlinear characteristic of some actuators and sensors is known as deadband. The deadband is typically a region of input close to zero at which the output remains zero. Once the input travels outside the deadband, then the output varies with input, as shown in Fig. 18.10. Analog joystick inputs frequently use a small amount of deadband to reduce the effect of noise from human inputs. A very small movement of the joystick produces no output, but the joystick acts normally with larger inputs.

Deadband is also commonly found in household thermostats and other process type controllers, as shown in Fig. 18.11. When a room warms and the temperature reaches the setpoint (or desired value)

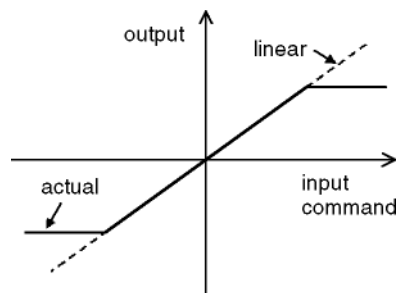



FIGURE 18.9 Saturation.

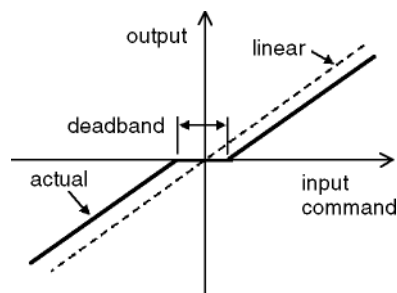



FIGURE 18.10 Deadband.

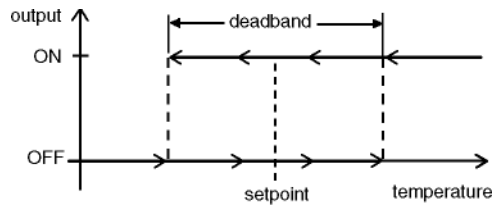



FIGURE 18.11 Thermostat deadband.

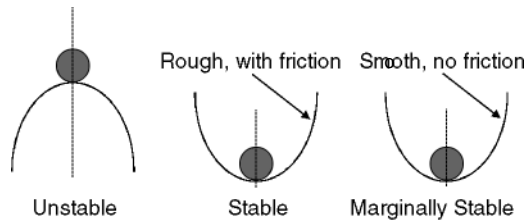



FIGURE 18.12 System stability.

on the thermostat, the output remains off. Once room temperature has increased to the setpoint plus half the deadband, then the cooling system output goes to fully on. As the room cools, the output stays fully on until the temperature reaches the setpoint minus half the deadband. At this point the cooling system output goes fully off.

18.14 System Response

Sensors and actuators respond to inputs that change with time. Any system that changes with time is considered a dynamic system. Understanding the response of dynamic systems to different types of inputs is important in mechatronic system design. The most important concept in system response is stability. The term stability has many different definitions and uses, but the most common definition is related to equilibrium. A system in equilibrium will remain in the same state in the absence of external disturbances. A stable system will return to an equilibrium state if a "small" disturbance moves the system away from the initial state. An unstable system will not return to an equilibrium position, and frequently will move "far" from the initial state.

Figure 18.12 illustrates three stability conditions with a simple ball and hill system. In each case an equilibrium position is easily identified—either the top of the hill or the bottom of the valley. In the unstable case, a small motion of the ball away from the equilibrium position will cause the ball to move "far" away, as it rolls down the hill. In the stable case, a small movement of the ball away from the equilibrium position will eventually result in the ball returning, perhaps after a few oscillations. In the third case, the absence of friction causes the ball to oscillate continuously about the equilibrium position once a small movement has occurred. This special case is often known as marginal stability, since the system never quite returns to the equilibrium position.

Most sensors and actuators are inherently stable. However, the addition of active control systems can cause a system of stable devices to exhibit overall unstable behavior. Careful analysis and testing is required to ensure that a mechatronic system acts in a stable manner. The complex response of stable dynamic systems is frequently approximated by much simpler systems. Understanding both first-order and second-order system responses to either instantaneous (or step) changes in inputs or sinusoidal inputs will suffice for most situations.

18.15 First-Order System Response

First-order systems contain two primary elements: an energy storing element and an element which dissipates (or removes) energy. Typical first-order systems include resistor–capacitor filters and resistor–inductor networks (e.g., a coil of a stepper motor). Thermocouples and thermistors also form first-order systems, due to thermal capacitance and resistance. The differential equation describing the time response of a generic first-order system is

$$\frac{dy(t)}{dt} + \frac{1}{\tau} y(t) = f(t)$$

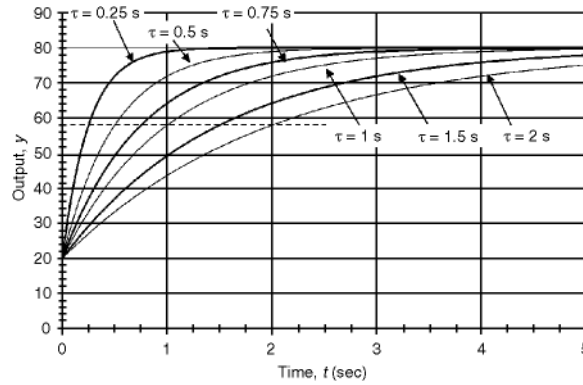



FIGURE 18.13 First-order system—step response.

where $y(t)$ is the dependent output variable (velocity, acceleration, temperature, voltage, etc.), t is the independent input variable (time), τ is the time constant (units of seconds), and $f(t)$ is the forcing function (or system input).

The solution to this equation for a step or constant input is given by

$$y(t) = y_\infty + (y_0 - y_\infty)e^{-t/t}$$

where y_∞ is the limiting or final (steady-state) value, y_0 is the initial value of the independent variable at $t = 0$.

A set of typical first-order system step responses is shown in Fig. 18.13. The initial value is arbitrarily selected as 20 with final values of 80. Time constants ranging from 0.25 to 2 s are shown. Each of these curves directly indicates its time constant at a key point on the curve. Substituting $t = \tau$ into the first-order response equation with $y_0 = 20$ and $y_\infty = 80$ gives

$$y(t) = 80 + (20 - 80)e^{-1} = 57.9$$

Each curve crosses the $y(\tau) \approx 57.9$ line when its time constant τ equals the time t. This concept is frequently used to experimentally determine time constants for first-order systems.

18.16 Underdamped Second-Order System Response

Second-order systems contain three primary elements: two energy storing elements and an element which dissipates (or removes) energy. The two energy storing elements must store different types of energy. A typical mechanical second-order system is the spring–mass–damper combination shown in Fig. 18.14. The spring stores potential energy ($PE = \frac{1}{2}kx^2$), while the mass stores kinetic energy ($KE = \frac{1}{2}mv^2$), where k is the spring stiffness (typical units of N/m), x is the spring deflection (typical units of m), m is the mass (typical units of kg), and v is the absolute velocity of the mass (typical units of m/s).

A common electrical second-order system is the resistor–inductor–capacitor (RLC) network, where the capacitor and inductor store electrical energy in two different forms. The generic form of the dynamic equation for an underdamped second-order system is

$$\frac{d^2y(t)}{dt^2} + 2\zeta\omega_n\frac{dy(t)}{dt} + \omega_n^2y(t) = f(t)$$

where $y(t)$ is the dependent variable (velocity, acceleration, temperature, voltage, etc.), t is the independent variable (time), ζ is the damping ratio (a dimensionless quantity), ω_n is the natural frequency (typical units of rad/s), and $f(t)$ is the forcing function (or input).

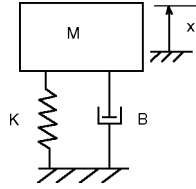



FIGURE 18.14 Spring–mass–damper system.

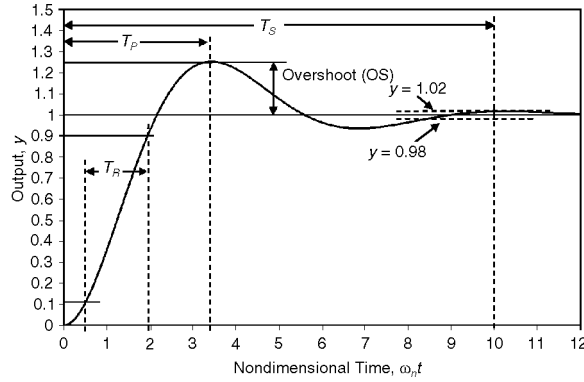



FIGURE 18.15 Second-order system—step response.

The response of an underdamped ($0 \le \zeta < 1$) second-order system to a *unit* step input can be determined as:

$$y(t) = 1 - e^{-\zeta\omega_n t}\left(\cos\omega_n\sqrt{1-\zeta^2}t + \frac{\zeta}{\sqrt{1-\zeta^2}}\sin\omega_n\sqrt{1-\zeta^2}t\right)$$

This second-order system step response is often characterized by a set of time response parameters illustrated in Fig. 18.15.

These time response parameters are functions of the damping ratio ζ and the natural frequency ω_n:

- peak time, T_P: the time required to reach the first (or maximum) peak

$$T_P = \frac{\pi}{\omega_n\sqrt{1-\zeta^2}}$$

- percent overshoot, %OS: amount the response exceeds or overshoots the steady-state value

$$\%OS = 100e^{-(\zeta\pi/\sqrt{1-\zeta^2})}$$

- settling time, T_S: the time when the system response remains within ±2% of the steady-state value

$$T_S = \frac{4}{\zeta\omega_n}$$

- rise time, T_R: time required for the response to go from 10% to 90% of the steady-state value. Figure 18.16 shows the nondimensional rise time ($\omega_n T_R$) as a function of damping ratio, *z*. A frequently used approximation relating these two parameters is

$$\omega_n T_R \approx 2.16\zeta + 0.6 \quad 0.3 \le \zeta 0.8$$

Figures 18.17 and 18.18 show the unit step response of a second-order system as a function of damping ratio ζ.

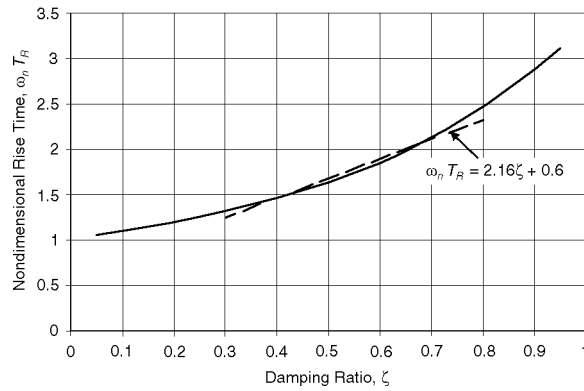



FIGURE 18.16 Rise time vs. damping ratio, ζ.

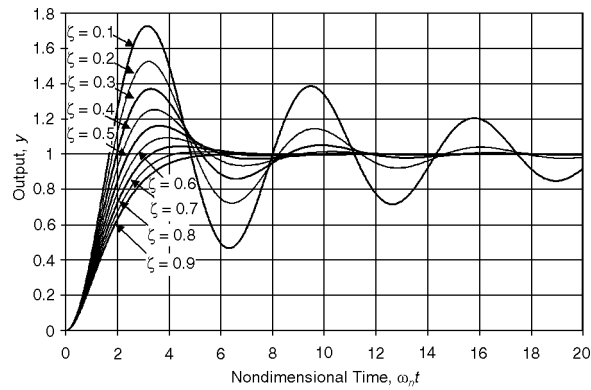



FIGURE 18.17 Second-order system step response vs. damping ratio, ζ.

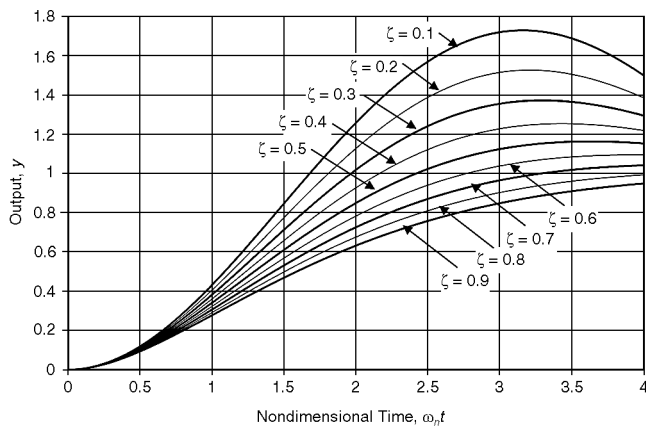



FIGURE 18.18 Initial second-order system step response vs. damping ratio, ζ.

18.17 Frequency Response

The response of any dynamic system to a sinusoidal input is called the frequency response. A generic first-order system with a sinusoidal input of amplitude A would have the dynamic equation of

$$\frac{dy(t)}{dt} + \frac{1}{\tau}y(t) = f(t) = A\sin(\omega t)$$

where ω is the frequency of the sinusoidal input and τ is the first-order time constant. The steady-state solution to this equation is

$$y(t) = AM\sin(\omega t + \Phi)$$

where $M = 1/\sqrt{(\tau\omega)^2 + 1}$ is the amplitude ratio (a dimensionless quantity), and $\Phi = -\tan^{-1}(\tau\omega)$ is the phase angle.

Figure 18.19 is a plot of the magnitude ratio M_{dB} and the phase angle Φ as a function of the non-dimensional frequency, $\tau\omega$. Note that the magnitude is frequently plotted in terms of decibels, where $M_{dB} = 20\log_{10}(M)$.

The frequency at which the magnitude ratio equals 0.707 (or –3 dB) is called the bandwidth. For a first-order system, the bandwidth is inversely proportional to the time constant. So, $\omega = 1/\tau$.

A generic second-order system with a sinusoidal input of amplitude A and frequency ω would have the dynamic equation of

$$\frac{d^2y(t)}{dt^2} + 2\zeta\omega_n\frac{dy(t)}{dt} + \omega_n^2 y(t) = A\sin(\omega t)$$

The steady-state solution to this equation is

$$y(t) = \frac{AM}{\omega_n^2}\sin(\omega t + \Phi)$$

where

$$M = \frac{1}{\sqrt{[1-(\omega^2/\omega_n^2)]^2 + [2\zeta(\omega/\omega_n)]^2}}$$

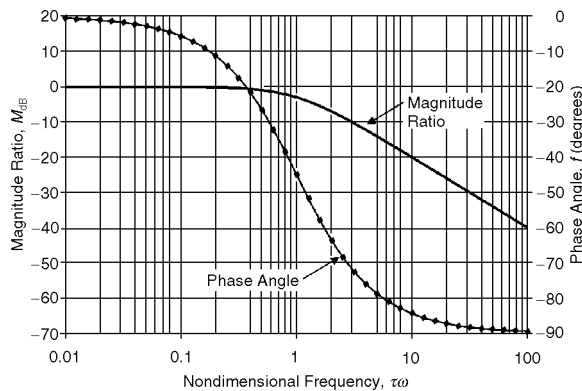



FIGURE 18.19 Frequency response for first-order system.

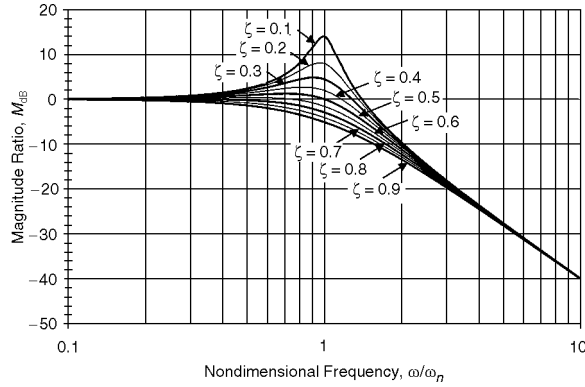



FIGURE 18.20 Frequency response magnitude for second-order system.

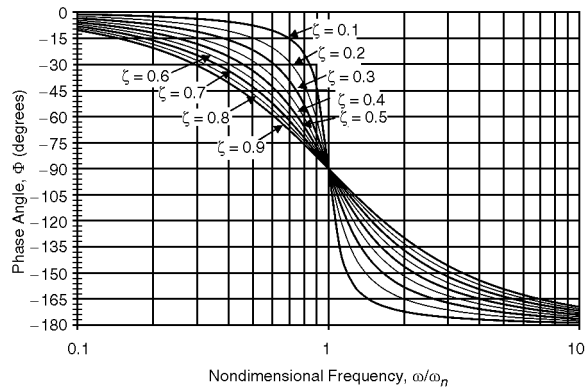



FIGURE 18.21 Frequency response phase angle for second-order system.

is the amplitude ratio (a dimensionless quantity), and

$$\Phi = -\tan^{-1}\left[\frac{2\zeta(\omega/\omega_n)}{1-(\omega^2/\omega_n^2)}\right]$$

is the phase angle.

Figures 18.20 and 18.21 are plots of the magnitude response $M_{dB} = 20\log_{10}(M)$ and the phase angle Φ for the second-order system as a function of damping ratio, ζ. The peak value in the magnitude response, M_P, can be found by taking the derivative of M with respect to ω and setting the result to zero to find (Nise, 1995)

$$M_P = \frac{1}{2\zeta\sqrt{1-\zeta^2}}$$

This peak value in M occurs at the frequency ω_P given by

$$\omega_P = \omega_n\sqrt{1-2\zeta^2}$$

The peak value in an experimentally determined frequency response can be used to estimate both the natural frequency and damping ratio for a second-order system. These parameters can then be used to estimate time domain responses such as peak time and percent overshoot.

Reference

Nise, N. S., *Control Systems Engineering,* 2nd ed., Benjamin/Cummings, 1995.

19
Sensors

Kevin M. Lynch
Northwestern University

Michael A. Peshkin
Northwestern University

Halit Eren
Curtin University of Technology

M. A. Elbestawi
McMaster University

Ivan J. Garshelis
Magnova, Inc.

Richard Thorn
University of Derby

Pamela M. Norris
University of Virginia

Bouvard Hosticka
University of Virginia

Jorge Fernando Figueroa
NASA Stennis Space Center

H. R. (Bart) Everett
Space and Naval Warfare Systems Center

Stanley S. Ipson
University of Bradford

Chang Liu
University of Illinois

19.1 Linear and Rotational Sensors

Kevin M. Lynch and Michael A. Peshkin

By far the most common motions in mechanical systems are linear translation along a fixed axis and angular rotation about a fixed axis. More complex motions are usually accomplished by composing these simpler motions. In this chapter we provide a summary of some of the many technologies available for sensing linear and rotational motion along a single axis. We have arranged the sensing modalities according to the physical effect exploited to provide the measurement.

Contact

The simplest kind of displacement sensor is a mechanical switch which returns one bit of information: touching or not touching. A typical *microswitch* consists of a lever which, when depressed, creates a mechanical contact within the switch, which closes an electrical connection (Fig. 19.1). Microswitches may be used as bump sensors for mobile robots, often by attaching a compliant material to the lever (such as a whisker) to protect the robot body from impact with a rigid obstacle. Another popular application of the microswitch in robotics is as a *limit switch*, indicating that a joint has reached the limit of its allowable travel.

Figure 19.2 shows a typical configuration for a microswitch. The pull-up resistor keeps the signal at +V until the switch closes, sending the signal to ground. As the switch closes, however, a series of micro-impacts may lead to "bounce" in the signal. A "debouncing" circuit may be necessary to clean up the output signal.

Switches may be designated NO or NC for normally open or normally closed, where "normally" indicates the unactivated or unpressed state of the switch. A switch may also have multiple *poles* (P) and one or two *throws* (T) for each pole. A pole moves as the switch is activated, and the throws are the possible contact points for the pole. Thus an SPDT (single pole double throw) switch switches a single pole from contact with one throw to the other, and a DPST (double pole single throw) switch switches two poles from open to closed circuit (Fig. 19.3).

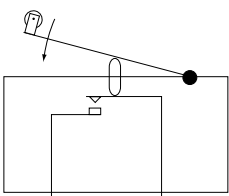



FIGURE 19.1 A typical microswitch.

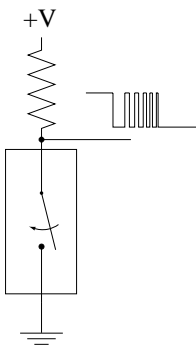



FIGURE 19.2 Signal bounce at a closing switch.

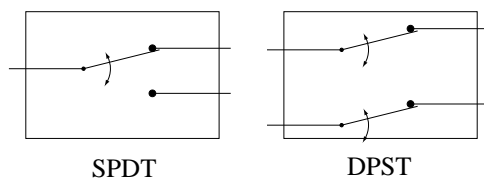



FIGURE 19.3 SPDT and DPST switch configurations.

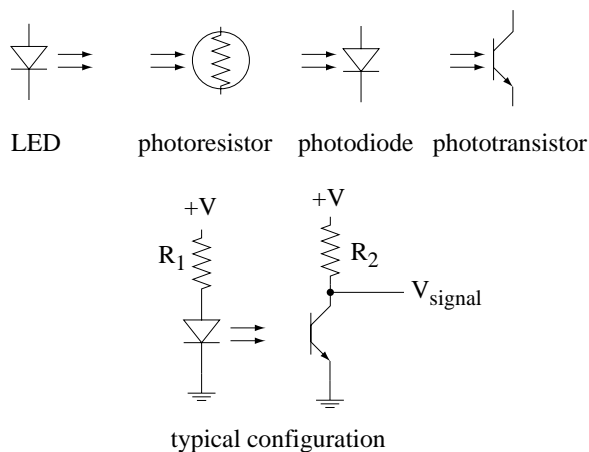



FIGURE 19.4 Optoelectronic circuit symbols and a typical emitter/detector configuration.

Infrared

Infrared light can be used in a variety of ways to measure linear and rotational displacement. Typically, an infrared light-emitting diode (LED), or *photoemitter*, is used as a source, and an infrared sensitive device is used to detect the emitted light. The detector could be a *photoresistor* or *photocell*, a variable resistor which changes resistance depending on the strength of the incident light (possibly infrared or visible); a *photodiode*, which allows the flow of electrical current in one direction in the presence of infrared light, and otherwise acts as an open circuit; or a *phototransistor*. In a phototransistor, the incident infrared light acts as the base current for the transistor, allowing the flow of collector current proportional to the strength of the received infrared light (up to saturation of the transistor). Circuit symbols for the various elements are shown in Fig. 19.4.

If the emitter and detector are facing each other, they can be used as a beam-breaker, to detect if something passes between. This is called a *photointerrupter* (Fig. 19.5). If the emitter and detector are free to move along the line connecting them, the strength of the received signal can be used to measure the distance separating them. Infrared photodetectors may be sensitive to ambient light, however. To distinguish the photoemitter light from background light, the source can be modulated (i.e., switched on and off at a high frequency), and the detector circuitry designed to respond only to the modulated infrared.

An emitter and detector facing the same direction can be used to roughly measure the distance to a nearby surface by the strength of the returned light reflecting off the surface. This is called a *photoreflector* (Fig. 19.6). Alternatively, such a sensor could be used to detect light absorbing or light reflecting surfaces at a constant distance, as in mobile robot line following. Light polarizing filters can also be used on the emitter and detector so that the detector only recognizes light reflected by a special “optically active” retroreflecting surface.

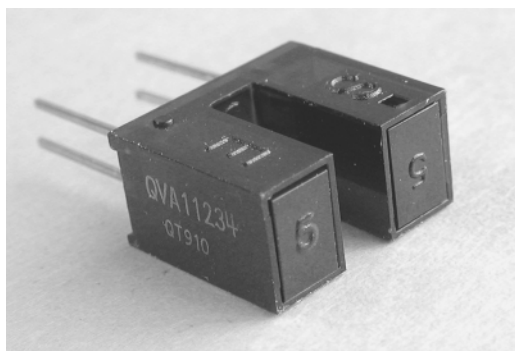



FIGURE 19.5 The Fairchild semiconductor QVA11234 photointerrupter.

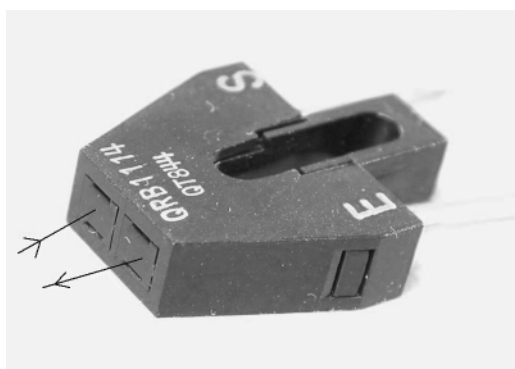



FIGURE 19.6 The Fairchild semiconductor QRB1114 photoreflective sensor.

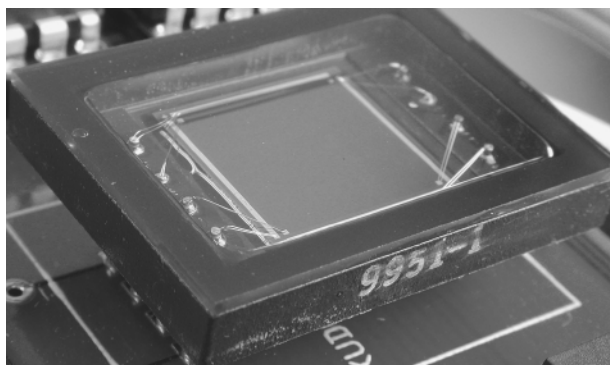



FIGURE 19.7 A position sensitive detector (PSD), UDT Sensors, Inc.

Photointerrupters and photoreflectors can be bought prepackaged or constructed separately from an infrared LED and a photodiode or phototransistor, after making certain the detector is sensitive to the wavelength produced by the LED.

Photoreflector units are also available with more advanced position sensitive detectors (PSDs), which report the location of infrared light incident on the sensing surface (Fig. 19.7). The fixed location of the LED relative to the PSD, as well as the location of the image of the infrared light on the PSD, allows the

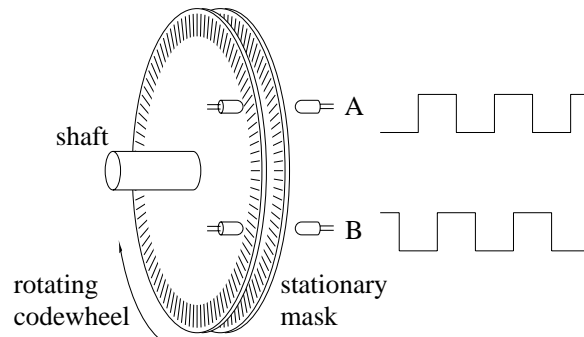



FIGURE 19.8 Schematic of an incremental encoder.

use of triangulation to determine the distance to the target. Such distance measuring sensors are manufactured by Sharp and Hamamatsu.

Optical Encoders

An optical encoder uses photointerrupters to convert motion into an electrical pulse train. These electrical pulses "encode" the motion, and the pulses are counted or "decoded" by circuitry to produce the displacement measurement. The motion may be either linear or rotational, but we focus on more common rotary optical encoders.

There are two basic configurations for rotary optical encoders, the *incremental* encoder and the *absolute* encoder. In an incremental encoder, a disk (or codewheel) attached to a rotating shaft spins between two photointerrupters (Fig. 19.8). The disk has a radial pattern of lines, deposited on a clear plastic or glass disk or cut out of an opaque disk, so that as the disk spins, the radial lines alternately pass and block the infrared light to the photodetectors. (Typically there is also a stationary mask, with the same pattern as the rotating codewheel, in the light path from the emitters to the detectors.) This results in pulse trains from each of the photodetectors at a frequency proportional to the angular velocity of the disk. These signals are labeled A and B, and they are 1/4 cycle out of phase with each other. The signals may come from photointerrupters aligned with two separate tracks of lines at different radii on the disk, or they may be generated by the same track, with the photointerrupters placed relative to each other to give out of phase pulse trains.

By counting the number of pulses and knowing the number of radial lines in the disk, the rotation of the shaft can be measured. The direction of rotation is determined by the phase relationship of the A and B pulse trains, i.e., which signal leads the other. For example, a rising edge of A while B = 1 may indicate counterclockwise rotation, while a rising edge of A while B = 0 indicates clockwise rotation. The two out-of-phase signals are known as quadrature signals.

Incremental encoders commonly have a third output signal called the index signal, labeled I or Z. The index signal is derived from a separate track yielding a single pulse per revolution of the disk, providing a home signal for absolute orientation. In practice, multiple photointerrupters can be replaced by a single source and a single array detecting device.

IC decoder chips are available to decode the pulse trains. The inputs to the chip are the A and B signals, and the outputs are one or more pulse trains to be fed into a counter chip. For example, the US Digital LS7083 outputs two pulse trains, one each for clockwise and counterclockwise rotation, which can be sent to the inputs of a 74193 counter chip (Fig. 19.9). Standard decoding methods for the quadrature input are 1X, 2X, and 4X resolution. In 1X resolution, a single count is generated for the rising or falling edge of just one of the pulse trains, so that the total number of encoder counts for a single revolution of the disk is equal to the number of lines in the disk. In 4X resolution, a count is generated for each rising and falling edge of both pulse trains, resulting in four times the angular resolution. An encoder

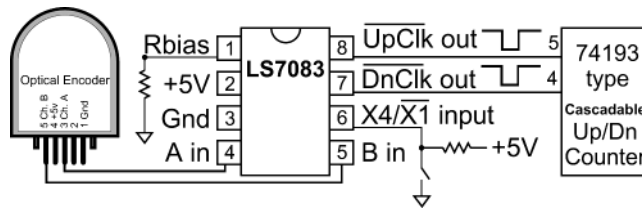



FIGURE 19.9 An optical encoder, US Digital LS7083 quadrature decoder chip, and counter (courtesy of US Digital, Inc.).

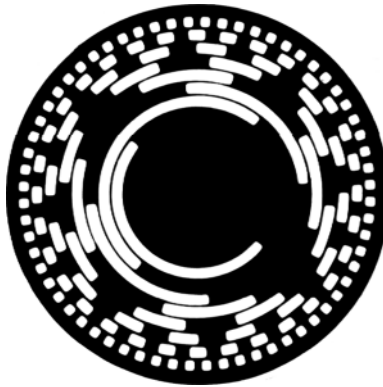



FIGURE 19.10 An 8-bit Gray code absolute encoder disk, courtesy of BEI Technologies Industrial Encoder Division.

with 1000 lines on the code wheel being decoded at 4X resolution yields an angular resolution of 360∞/ $(4 \times 1000) = 0.09°$.

While a *single-ended output* encoder provides the signals A, B, and possibly Z, a *differential output* encoder also provides the complementary outputs A′, B′, and Z′. Differential outputs, when used with a differential receiver, can increase the electrical noise immunity of the encoder.

A drawback of the incremental encoder is that there is no way to know the absolute position of the shaft at power-up without rotating it until the index pulse is received. Also, if pulses are momentarily garbled due to electrical noise, the estimate of the shaft rotation is lost until the index pulse is received. A solution to these problems is the absolute encoder. An absolute encoder uses k photointerrupters and k code tracks to produce a k-bit binary word uniquely representing 2^k different orientations of the disk, giving an angular resolution of $360°/2^k$ (Fig. 19.10). Unlike an incremental encoder, an absolute encoder always reports the absolute angle of the encoder.

The radial patterns on the tracks are arranged so that as the encoder rotates in one direction, the binary word increments or decrements according to a binary code. Although natural binary code is a possibility, the Gray code is a more common solution. With natural binary code, incrementing by one may change many or all of the bits, e.g., 7 to 8 in decimal is 0111 to 1000 in natural binary. With the Gray code, only one bit changes as the number increments or decrements, e.g., 7 to 8 in decimal is 0100 to 1100 in Gray code. The rotational uncertainty during a Gray code transition is only a single count, or $360°/2^k$. With the natural binary code, an infinitesimal misalignment between the lines and the photointerrupters may cause the reading to briefly go from 0111 (7) to 1111 (15) during the transition to 1000 (8).

In general, incremental encoders provide higher resolution at a lower cost and are the most common choice for many industrial and robotic applications.

Resistive

One of the simplest and least expensive ways to measure rotational or linear motion is using a variable resistor called a *potentiometer* or *rheostat*. We focus on rotary potentiometers, or "pots" for short, but the principle of operation is the same in the linear case.

A pot consists of three terminals (Fig. 19.11(a,b)). Two end terminals, call them terminals 1 and 3, connect to either end of a length of resistive material, such as partially conductive plastic, ceramic, or a long thin wire. (For compactness, the long wire is wound around in loops to make a coil, leading to the name *wirewound* potentiometer.)

The other terminal, terminal 2, is connected to a *wiper*, which slides over the material as the pot shaft rotates. The total resistance of the pot R_{13} is equal to the sum of the resistance R_{12} between terminal 1 and the wiper, and the resistance R_{23} between the wiper and terminal 3. Typically the wiper can rotate from one end of the resistive material ($R_{13} = R_{12}$) to the other ($R_{13} = R_{23}$). If the full motion of the wiper is caused by one revolution of the shaft or less, the pot is called a *single-turn* pot. If the full motion is caused by multiple revolutions, it is called a *multi-turn* pot.

Typically a pot is used by connecting terminal 1 to a voltage V, terminal 3 to ground, and using the voltage at the wiper as a measure of the rotation. The voltage observed at the wiper is $V(R_{23}/R_{13})$ and is a linear function of the rotation of the shaft.

A remarkably simple absolute sensor for a wide range of distances is the string pot or draw-wire sensor (Fig. 19.12). It consists of a string wrapped on a spool, with a potentiometer to monitor rotations of the spool. A return spring keeps the string taut. Lengths up to many meters may be measured, using sensors incorporating multi-turn pots. The same technique is similarly useful for short distances (a few centimeters) using compact single-turn pots and a small spool. Both tolerate misalignment or arc-like motion well. String pots are susceptible to damage to the string in exposed applications, but the sensor element is small and unobtrusive. Manufacturers include RDP Electronics, SpaceAge Control, and UniMeasure.

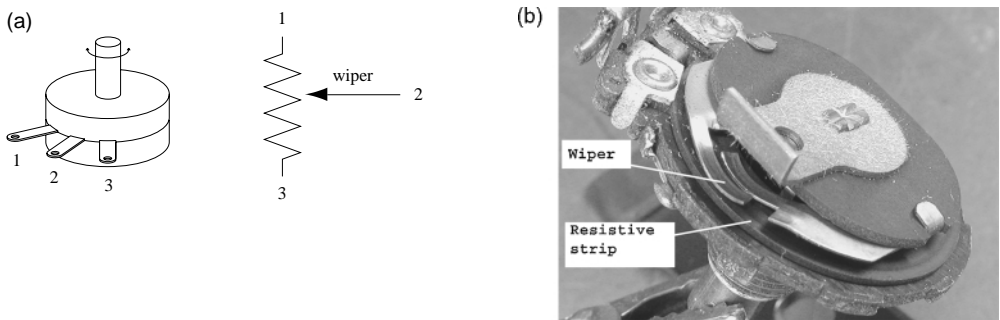



FIGURE 19.11 (a) As the shaft of the potentiometer rotates, the wiper moves from one end of the resistive material to the other. (b) The inside of a typical potentiometer, showing the wiper contacting a resistive strip.

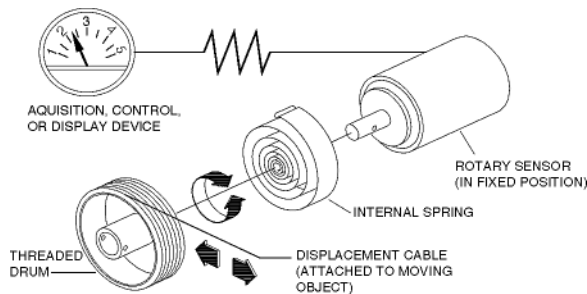



FIGURE 19.12 A string pot, courtesy of Space Age Control, Inc.

Another type of resistive sensor is the flexible bend sensor. Conductive ink between two electrical contacts on a flexible material changes resistance as the material bends and stretches. Used in a voltage divider with a fixed resistor, the analog voltage may be used as a measure of the bend. Such a sensor could be used to detect contact (like a whisker) or as a rough measure of the deformation of a surface to which it is attached.

Tilt (Gravity)

A *mercury switch* can be used to provide one bit of information about orientation relative to the gravity vector. A small drop of mercury enclosed in a glass bulb opens or closes the electrical connection between two leads depending on the orientation of the sensor. Several mercury switches at different orientations may be used to get a rough estimate of tilt. The signal from a mercury switch may "bounce" much like the signal from a mechanical contact switch (Fig. 19.2).

An *inclinometer* can be used to measure the amount of tilt. One example is the *electrolytic tilt sensor*. Manufacturers include The Fredericks Company and Spectron Glass. Two-axis models have five parallel rod-like electrodes in a sealed capsule, partially filled with a conductive liquid. Four of the electrodes are at the corners of a square, with one in the middle. Tilting the sensor changes the distribution of current injected via the center electrode in favor of the electrodes which are more deeply immersed.

Tilt sensors may be obtained with liquids of varying viscosity, to minimize sloshing. Because a DC current through the liquid would cause electrolysis and eventually destroy the sensor, AC measurements of conductivity are used. As a result, the support electronics are not trivial.

The liquid conductivity is highly temperature dependent. The support electronics for the tilt sensor must rely on a ratio of conductivity between pairs of rods. Also, although the electrolytic tilt sensor operates over a wide temperature range, it is greatly disturbed by nonuniformities of temperature across the cell.

Another kind of simple inclinometer can be constructed from a rotary potentiometer with a pendulum bob attached. A problem with this solution is that friction may stop the bob's motion when it is not vertical. A related idea is to use an absolute optical encoder with a pendulum bob. Complete sensors operating on this principle can be purchased with advanced options, such as magnetic damping to reduce overshoot and oscillation. An example is US Digital's 12-bit A2I absolute inclinometer.

Of course, gravity acting on a device is indistinguishable from acceleration. If the steady-state tilt of a device is the measurement of interest, simple signal conditioning should be used to ensure that the readings have settled.

Other more sophisticated tilt sensors include gyroscopes and microelectromechanical (MEMS) devices, which are not discussed here.

Capacitive

Capacitance can be used to measure proximity or linear motions on the order of millimeters. The capacitance C of a parallel plate capacitor is given by $C = \varepsilon_r \varepsilon_o A/d$, where ε_r is the relative permittivity of the dielectric between the plates, ε_o is the permittivity of free space, A is the area of overlap of the two plates, and d is the plate separation. As the plates translate in the direction normal to their planes, C is a nonlinear function of the distance d. As the plates translate relative to each other in their planes, C is a linear function of the area of overlap A. Used as proximity sensors, capacitive sensors can detect metallic or nonmetallic objects, liquids, or any object with a dielectric constant greater than air.

One common sensing configuration has one plate of the capacitor inside a probe, sealed in an insulator. The external target object forms the other plate of the capacitor, and it must be grounded to the proximity sensor ground. As the sensor approaches the target, the capacitance increases, modifying the oscillation of a detector circuit including the capacitor. This altered oscillation may be used to signal proximity or to obtain a distance measurement.

Manufacturers of capacitive sensors include Cutler-Hammer and RDP Electronics.

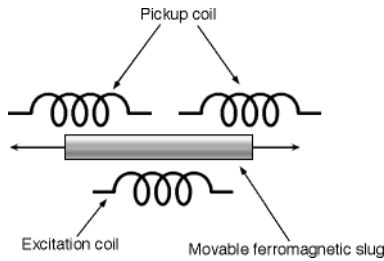



FIGURE 19.13 Operating principle of an LVDT.

AC Inductive

LVDT

The best known AC inductive sensor is the *linear variable differential transformer*, or LVDT. The LVDT is a tube with a plunger, the displacement of the plunger being the variable to be measured (Fig. 19.13). The tube is wrapped with at least two coils, an excitation coil and a pickup coil. An AC current (typically 1 kHz) is passed through the excitation coil, and an AC signal is detected from the pickup coil and compared in magnitude and in phase (0 or 180°) to the excitation current. Support electronics are needed for the demodulation, which is called synchronous detection. The plunger carries a ferromagnetic slug, which enhances the magnetic coupling from the excitation coil to the pickup coil. Depending on the position of the slug within the pickup coil, the detected signal may be zero (when the ferrite slug is centered in the pickup coil), or increasing in amplitude in one or the other phase, depending on displacement of the slug.

LVDTs are a highly evolved technology and can be very accurate, in some cases to the micron level. They have displacement ranges of millimeters up to a meter. They do not tolerate misalignment or nonlinear motion, as a string pot does.

Resolvers

A *resolver* provides a measure of shaft angle, typically with sine and cosine analog outputs. It uses an AC magnetic technique similar to the LVDT, and similar support electronics to provide synchronous detection. Resolvers are very rugged and for this reason are often preferred over optical encoders on motor shafts, although they are not as accurate and they have greater support electronics requirements. Some resolver drives have extra outputs as if they were incremental encoders, for compatibility. Additionally, resolvers provide an absolute measure of shaft angle. The resolver, like the LVDT, is a well established and evolved technology.

DC Magnetic

A magnetic field acting on moving electrons (e.g., a current in a semiconductor) produces a sideways force on the electrons, and this force can be detected as a voltage perpendicular to the current. The effect is small, even in semiconductors, but has become the basis of a class of very rugged, inexpensive, and versatile sensors.

Hall Effect Switches

Hall effect switches refers to devices which produce a binary output, depending on whether the magnetic field intensity exceeds a threshold or not. In their component form, these switches may be packaged as 3-terminal devices the size of a transistor package (TO-92) or surface mount, having only a power lead (3–24 V), a ground lead, and an output lead. Typically the output is pulled to ground, or not, depending on the magnetic state. Hall effect switches are also available in environmental packages of all sorts.

The actuation threshold ranges from a few gauss (the Earth's magnetic field is 1/2 G) up to the hundreds of gauss levels typical of permanent magnets. Often there is a fair degree of unit-to-unit variability in threshold.

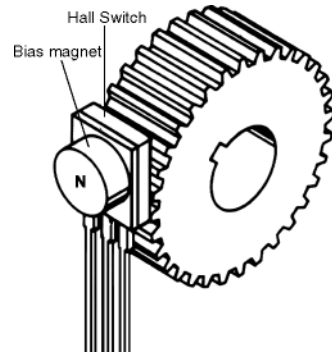



FIGURE 19.14 Detecting gear teeth in a ferrous material using a Hall switch and a bias magnet. Courtesy of Allegro Microsystems, Inc.

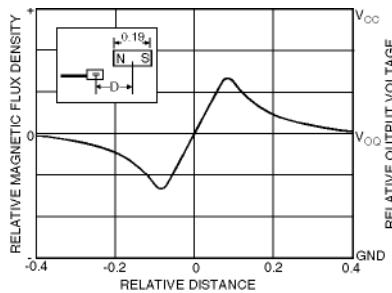



FIGURE 19.15 Output of an analog Hall sensor vs. position relative to a magnet. Courtesy of Allegro Microsystems, Inc.

Hall effect switches are hysteretic: their "turn-on" threshold may be different than the "turn-off" threshold. Sometimes hysteresis is used to make a switch latching, so that it stays in its last state (on or off) until the applied magnetic field is reversed. Non-latching Hall switches may be unipolar (responding only to one orientation of magnetic field) or bipolar (responding to a field of either polarity). Turn-on and turn-off times are in microseconds.

Hall switches have wide operating temperature ranges and are often used in automobile engine compartments. Another advantage is that they are not susceptible to most of the fouling mechanisms of optical or mechanical switches, such as liquids or dirt. While often the moving part that is detected is a magnet, it can also be arranged that a stationary "bias" magnet is intensified in its effect on the hall switch by the approach of a ferrous part, such as a gear tooth, thus allowing nonmagnetized objects to be detected (Fig. 19.14).

Typical applications are the detection of a moving part, replacing a mechanical limit switch. The Hall switch has no moving or exposed parts and is wear-free. Another common use is in indexing of rotational or translational motion. The Hall switch is installed to detect one position, and its output pulse is used as a reference for an incremental encoder which can count distance from that reference point. Hall switches are inexpensive and small, so a number of them can be spaced at intervals of millimeters, forming a low-resolution linear or rotational encoder or multi-position switch. Such an encoder or switch has the ruggedness advantages of Hall switches.

Analog Hall Sensors

In a package the same small size as Hall switches, and costing little more, one can also get Hall devices that have an analog output proportional to magnetic field strength (Fig. 19.15). Typically, these have full-scale magnetic field sensitivity in the 100 G range. These are not useful as a compass in the Earth's sub-gauss magnetic field.

Hall sensors are useful as linear or rotational encoders. Two Hall sensors may be arranged at right angles to detect the sine and cosine of the angle of a rotating magnet, thus forming an absolute rotation sensor.

Commercially available devices of this nature are called "Hall potentiometers" and have a variety of outputs (e.g., sine and cosine, or a linear ramp repeating with each revolution). In contrast to potentiometers with resistive strips and sliders, Hall pots allow continuous 360∞ rotation and experience no wear. All Hall effect devices are susceptible to external magnetic fields, however.

Hall sensors are also excellent transducers of short linear or arc-like motions. The motion of a bar magnet past a Hall sensor exposes the sensor to a magnetic field—which can be arranged to vary linearly with displacement—over a range of several millimeters up to several centimeters. (The bar magnet travels less than its own length.) Commercial implementations are known as throttle position sensors.

Tape-Based Sensors

There are a number of linear and rotational sensors, both incremental and absolute, which are similar to optical encoders but use magnetic patterns rather than optical ones. Linear applications are likely to require an exposed strip. In exposed applications, magnetic sensors have advantages in resistance to dirt, although the magnetic stripes must be protected from damage.

Ultrasonic

Ultrasonic (US) sensors use the time-of-flight of a pulse of ultrasonic sound through air or liquid to measure distance. Sensors are available with ranges from a few centimeters to 10 m. A great advantage of US sensors is that all of the sensor's electronic and transducer components are in one location, out of harm's way. The corresponding disadvantages are that US sensors tend to be indiscriminate: they may detect spurious targets, even very small ones, especially if these are near the transducer. Sensors are available with carefully shaped beams (down to 7∞) to minimize detection of spurious targets. Some include compensation for variation in air temperature, which affects sound velocity. US sensors can be used in surprising geometries. For instance, they can be used to detect the liquid level in a vertical pipe; back-reflection of sound pulses from the walls of a smooth pipe are minimal.

There is also an inexpensive and easily interfaced US sensor from Polaroid, derived from a ranging device for an instant camera, which is popular with experimenters.

Ultrasonic sensors typically have an analog output proportional to distance to target. Accuracies of the 1% level can be expected in a well-controlled environment.

Magnetostrictive Time-of-Flight

A more accurate technique for using time-of-flight to infer distance is the *magnetostrictive wire transducer* (MTS). A moving magnet forms the "target" corresponding to the acoustic target in US sensors, and need not touch the magnetrostrictive wire which is the heart of the device. The magnet's field acting on the magnetrostrictive wire creates an ultrasonic pulse in the wire when a current pulse is passed through the wire. The time interval from the current pulse to the detection of the ultrasonic pulse at the end of the wire is used to determine the position of the magnet along the wire (Fig. 19.16).

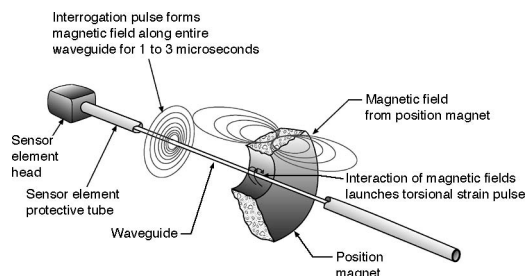



FIGURE 19.16 Principle of operation of a magnetostrictive linear position sensor, courtesy of Temposonics, Inc.

The magnetostrictive transducer does not have the inherent compactness and ruggedness of ultrasonic through air, but does achieve similarly large measurement lengths, up to several meters. Accuracy and stability are excellent, far better than ultrasonic. Some misalignment or nonlinear motion is tolerated, because the target magnet does not need to be in very close proximity to the magnetostrictive wire.

Laser Interferometry

Laser interferometers are capable of measuring incremental linear motions with resolution on the order of nanometers. In an interferometer, collimated laser light passes through a beam-splitter, sending the light energy on two different paths. One path is directly reflected to the detector, such as an optical sensing array, giving a flight path of fixed length. The other path reflects back to the detector from a retroreflector (mirror) attached to the target to be measured. The two beams constructively or destructively interfere with each other at the detector, creating a pattern of light and dark fringes. The interference pattern can be interpreted to find the phase relationship between the two beams, which depends on the relative lengths of the two paths, and therefore the distance to the moving target. As the target moves, the pattern repeats when the length of the variable path changes by the wavelength of the laser. Thus the laser interferometer is inherently an incremental measuring device.

Laser interferometers are easily the most expensive sensors discussed in this chapter. They also have the highest resolution. Laser interferometers are very sensitive to mechanical misalignment and vibrations.

More information about sensors may be found in Sensors magazine (http://www.sensorsmag.com/).

References

1. Histand, M. B. and Alciatore, D. G., *Introduction to Mechatronics and Measurement Systems,* McGraw-Hill, Boston, MA, 1999.
2. Bolton, W., *Mechatronics,* 2nd edition, Addison Wesley Longman, New York, NY, 1999.
3. Horowitz, P. and Hill, W., *The Art of Electronics,* 2nd edition, Cambridge University Press, Cambridge, UK, 1998.
4. Auslander, D. M. and Kempf, C. J., *Mechatronics: Mechanical System Interfacing,* Prentice-Hall, Upper Saddle River, NJ, 1996.
5. Jones, J. L., Flynn, A. M., and Seiger, B. A., *Mobile Robots: Inspiration to Implementation,* 2nd edition, A. K. Peters, Boston, MA, 1999.

19.2 Acceleration Sensors

Halit Eren

Acceleration relating to motion is an important section of kinematic quantities: position, velocity, acceleration, and jerk. Each one of these quantities has a linear relationship with the neighboring ones. That is, all the kinematic quantities can be derived from a single quantity. For example, acceleration can be obtained by differentiating the corresponding velocity or by integrating the jerk. Likewise, velocity can be obtained by differentiating the position or by integrating the acceleration. In practice, only integration is widely used since it provides better noise characteristics and attenuation.

There are two classes of acceleration measurements techniques: *direct* measurements by specific accelerometers and *indirect* measurements where velocity is differentiated. The applicability of these techniques depends on the type of motion (rectilinear, angular, or curvilinear motion) or equilibrium centered vibration. For rectilinear and curvilinear motions, the direct measurement accelerometers are preferred. However, the angular acceleration is usually measured by indirect methods.

Acceleration is an important parameter for general-purpose absolute motion measurements, vibration, and shock sensing. For these measurements, accelerometers are commercially available in a wide range and many different types to meet diverse application requirements, mainly in three areas: (1) *Commercial applications*—automobiles, ships, appliances, sports and other hobbies; (2) *Industrial applications*—robotics, machine control, vibration testing and instrumentation; and (3) *High reliability applications*—military, space and aerospace, seismic monitoring, tilt, vibration and shock measurements.

Accelerometers have been in use for many years. Early accelerometers were mechanical types relying on analog electronics. Although early accelerometers still find many applications, modern accelerometers are essentially semiconductor devices within electronic chips integrated with the signal processing circuitry. Mechanical accelerometers detect the force imposed on a mass when acceleration occurs. A new type of accelerometer, the thermal type, senses the position through heat transfer.

Overview of Accelerometer Types

A basic accelerometer consists of a mass that is free to move along a sensitive axis within a case. The technology is largely based on this basic accelerometer and can be classified in a number of ways, such as mechanical or electrical, active or passive, deflection or null-balance accelerometers, etc. The majority of industrial accelerometers are classified as either deflection or null-balance types. Accelerometers used in vibration and shock measurements are usually the deflection types, whereas those used for the measurement of motions of vehicles, aircraft, and so on for navigation purposes may be either deflection or null-balance type.

This article will concentrate on the direct measurements of acceleration, which can be achieved by the accelerometers of the following types:

- Inertial and mechanical
- Electromechanical
- Piezoelectric
- Piezoresistive
- Strain gauges
- Capacitive and electrostatic force balance
- Micro- and nanoaccelerometers

Depending on the principles of operations, these accelerometers have their own subclasses.

Dynamics and Characteristics of Accelerometers

Acceleration is related to motion, a vector quantity, exhibiting a direction as well as magnitude. The direction of motion is described in terms of some arbitrary Cartesian or orthogonal coordinate systems. Typical rectilinear, angular, and curvilinear motions are illustrated in Figs. 19.17(a–c), respectively.

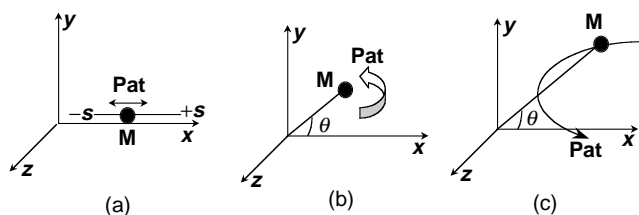



FIGURE 19.17 Types of motions to which accelerometers are commonly applied.

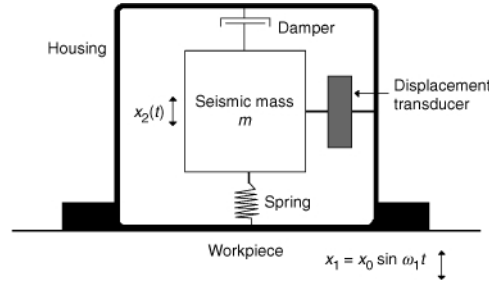



FIGURE 19.18 A typical seismic accelerometer.

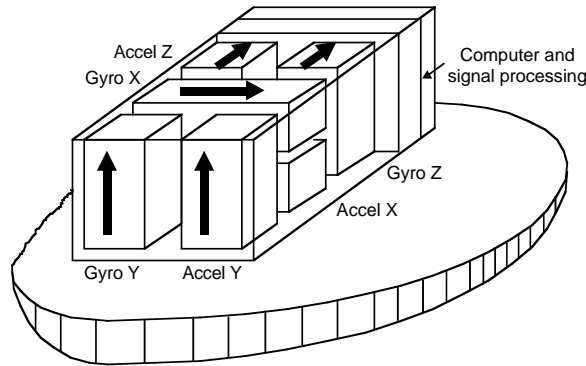



FIGURE 19.19 Arrangements of accelerometers in navigation and guidance systems.

The governing equations of these motions are as follows:

$$\text{Rectilinear acceleration} \quad a = \lim_{\Delta t \to 0} \frac{\Delta v}{\Delta t} = \frac{dv}{dt} = \frac{d(ds/dt)}{dt} = \frac{d^2 s}{dt^2} \quad (19.1)$$

$$\text{Angular acceleration} \quad \alpha = \lim_{\Delta t \to 0} \frac{\Delta \omega}{\Delta t} = \frac{d\omega}{dt} = \frac{d(d\theta/dt)}{dt} = \frac{d^2 \theta}{dt^2} \quad (19.2)$$

$$\text{Curvilinear acceleration} \quad a = \frac{dv}{dt} = \frac{d^2 x}{dt^2} i + \frac{d^2 y}{dt^2} j + \frac{d^2 z}{dt^2} k \quad (19.3)$$

where a and α are the accelerations; v and ω are the speeds; s is the distance; θ is the angle; i, j, and k are the unit vectors in x, y, and z directions, respectively.

For the correct applications of the accelerometers, a sound understanding of the characteristics of the motion under investigation is very important. The application areas may be linear and vibratory motion, angular motion, monitoring of the tilt of an object, or various forms of combinations. In each case, the correct selection and mounting of the accelerometer is necessary.

The majority of accelerometers can be viewed and analyzed as a single-degree-of-freedom seismic instrument that can be characterized by a mass, a spring, and a damper arrangement as shown in Fig. 19.18. In the case of multi-degrees-of-freedom systems, the principles of curvilinear motion can be applied as in Eq. (19.3) and multiple transducers must be used to create uniaxial, biaxial, or triaxial sensing points of the measurements. A typical example is the inertial navigation and guidance systems as illustrated in Fig. 19.19. In such applications, acceleration sensors play an important role in orientation and direction finding. Usually, miniature triaxial sensors detect changes in roll, pitch, and azimuth in x, y, and z directions.

If a single-degree-of-freedom system behaves linearly in a time invariant manner, the basic second-order differential equation describing the motion of the forced mass-spring system can be written as

$$f(t) = m\frac{d^2x}{dt^2} + c\frac{dx}{dt} + kx \tag{19.4}$$

where $f(t)$ is the force, m is the mass, c is the velocity constant, and k is the spring constant.

Nevertheless, the base of the accelerometer is in motion too. When the base is in motion, the force is transmitted through the spring to the suspended mass, depending on the transmissibility of the force to the mass. Equation (19.4) may be generalized by taking the effect motion of the base into account as

$$m\frac{d^2z}{dt^2} + c\frac{dz}{dt} + kz = mg\cos\theta - m\frac{d^2x_1}{dt^2} \tag{19.5}$$

where $z = x_2 - x_1$ is the relative motion between the mass and the base, x_1 is the displacement of the base, x_2 is the displacement of mass, and θ is the angle between the sense axis and gravity.

The complete solution to Eq. (19.5) can be obtained by applying the superposition principle. The superposition principle states that if there are simultaneously superimposed actions on a body, the total effect can be obtained by summing the effects of each individual action. Using superposition and using Laplace transforms gives

$$\frac{X(s)}{F(s)} = \frac{1}{ms^2} + cs + k \tag{19.6}$$

or

$$\frac{X(s)}{F(s)} = \frac{K}{s^2/\omega_n^2 + 2\zeta s/\omega_n + 1} \tag{19.7}$$

where s is the Laplace operator, $K = 1/k$ is the static sensitivity, $\omega_n = \sqrt{k/m}$ is the undamped critical frequency (rad/s), and $\zeta = (c/2)\sqrt{km}$ is the damping ratio.

As can be seen in the performance of accelerometers, the important parameters are the static sensitivity, the natural frequency, and the damping ratio, which are all functions of mass, velocity, and spring constants. Accelerometers are designed and manufactured to have different characteristics by suitable selection of these parameters. A short list of major manufacturers is given in Table 19.1.

Vibrations

This section is concerned with applications of accelerometers to measure physical properties such as acceleration, vibration and shock, and the motion in general. Although there may be fundamental differences in the types of motions, a sound understanding of the basic principles of the vibration will lead to the applications of accelerometers in different situations by making appropriate corrective measures.

Vibration is an oscillatory motion resulting from application of varying forces to a structure. The vibrations can be periodic, stationary random, nonstationary random, or transient.

Periodic Vibrations

In periodic vibrations, the motion of an object repeats itself in an oscillatory manner. This can be represented by a sinusoidal waveform

$$x(t) = X_p \sin(\omega t) \tag{19.8}$$

TABLE 19.1 List of Manufacturers

Analog Devices, Inc. 1 Technology Way, P.O. Box 9106 Norwood, MA 02062-9106 USA Tel: 781-329-4700 Fax: 781-326-8703	Kistler Instrument Corp. 75 John Glenn Dr. Amherst, NY 14228 2119 USA Tel: 888-KISTLER (547-8537) Fax: 716-691-5226
Aydin Telemetry 47 Friends Lane & Penns Trail Newtown, PA 18940 0328 USA Tel: 215-968-4271 Fax: 215-968-3214	Oceana Sensor Technologies, Inc. 1632-T Corporate Landing Pkwy. Virginia Beach, VA 23454 USA Tel: 757-426-3678 Fax: 757-426-3633
Bruel & Kjaer 2815-A Colonnades Court Norcross, GA 30071 USA Tel: 800-332-2040 Fax: 770-447-8440	PCB Piezotronics, Inc. 3425-T Walden Ave. Depew, NY 14043 2495 USA Tel: 716-684-0001 Fax: 716-684-0987
Dytran Instruments, Inc. 21592 Marilla St. Chatsworth, CA 91311 USA Tel: 800-899-7818 Fax: 800-899-7088	Piezo Systems, Inc. 186 Massachusetts Ave. Cambridge, MA 02139-4229 USA Tel: 617-547-1777 Fax: 617-354-2200
Endevco Corp. 30700 Rancho Viejo Rd. San Juan Capistrano, CA 92675 USA Tel: 800-982-6732 Fax: 949-661-7231	Rieker Instrument Co. P.O. Box 128 Folcroft, PA 19032 0128 USA Tel: 610-534-9000 Fax: 610-534-4670
Entran Devices, Inc. 10-T Washington Ave. Fairfield, NJ 07004 USA Tel: 888-8-ENTRAN (836-8726) Fax: 973-227-6865	Sensotec, Inc. 2080 Arlingate Lane Columbus, OH 43228 USA Tel: 800-858-6184 Fax: 614-850-1111
GS Sensors, Inc. 16 W. Chestnut St. Ephrata, PA 17522 USA Tel: 717-721-9727 Fax: 717-721-9859	Techkor Instrumentation 2001 Fulling Mill Rd. P.O. Box 70 Dept. T-1 New Cumberland, PA 17057-0070 USA Tel: 800-697-4567 Fax: 717-939-7170

where $x(t)$ is the time-dependent displacement, $\omega = 2\pi ft$ is the angular frequency, and X_p is the maximum displacement from a reference point.

The velocity of the object is the time rate of change of displacement,

$$v(t) = \frac{dx}{dt} = \omega X_p \cos(\omega t) = V_p \sin(\omega t + \pi/2) \qquad (19.9)$$

where $v(t)$ is the time-dependent velocity, and $V_p = \omega X_p$ is the maximum velocity.

The acceleration of the object is the time rate of change of velocity,

$$a(t) = \frac{dv}{dt} = \frac{d^2x}{dt^2} = -\omega^2 X_p \sin(\omega t) = A_p \sin(\omega t + \pi) \qquad (19.10)$$

where $a(t)$ is the time-dependent acceleration, and $A_p = \omega^2 X_p = \omega V_p$ is the maximum acceleration.

From the preceding equations it can be seen that the basic form and the period of vibration remains the same in acceleration, velocity, and displacement. But velocity leads displacement by a phase angle of 90∞ and acceleration leads velocity by another 90∞.

In nature, vibrations can be periodic, but not necessarily sinusoidal. If they are periodic but nonsinusoidal, they can be expressed as a combination of a number of pure sinusoidal curves, determined by Fourier analysis as

$$x(t) = X_0 + X_1 \sin(\omega_1 t + \Phi_1) + X_2 \sin(\omega_2 t + \Phi_2) + \cdots + X_n \sin(\omega_n t + \Phi_n) \tag{19.11}$$

where ω_1, ω_2,…, ω_n are the frequencies (rad/s), X_0, X_1,…, X_n are the maximum amplitudes of respective frequencies, and ϕ_1, ϕ_2,…, ϕ_n are the phase angles.

The number of terms may be infinite, and the higher the number of elements, the better the approximation. These elements constitute the frequency spectrum. The vibrations can be represented in the time domain or frequency domain, both of which are extremely useful in analysis.

Stationary Random Vibrations

Random vibrations are often met in nature, where they constitute irregular cycles of motion that never repeat themselves exactly. Theoretically, an infinitely long time record is necessary to obtain a complete description of these vibrations. However, statistical methods and probability theory can be used for the analysis by taking representative samples. Mathematical tools such as probability distributions, probability densities, frequency spectra, cross-correlations, auto-correlations, digital Fourier transforms (DFTs), fast Fourier transforms (FFTs), auto-spectral-analysis, root mean squared (rms) values, and digital filter analysis are some of the techniques that can be employed.

Nonstationary Random Vibrations

In this case, the statistical properties of vibrations vary in time. Methods such as time averaging and other statistical techniques can be employed.

Transients and Shocks

Often, short duration and sudden occurrence vibrations need to be measured. Shock and transient vibrations may be described in terms of force, acceleration, velocity, or displacement. As in the case of random transients and shocks, statistical methods and Fourier transforms are used in the analysis.

Typical Error Sources and Error Modeling

Acceleration measurement errors occur due to four primary reasons: sensors, acquisition electronics, signal processing, and application specific errors. In the direct acceleration measurements, the main error sources are the sensors and data acquisition electronics. These errors will be discussed in the biasing section and in some cases, sensor and acquisition electronic errors may be as high as 5%. Apart from these errors, sampling and A/D converters introduce the usual errors, which are inherent in them and exist in all computerized data acquisition systems. However, the errors may be minimized by the careful selection of multiplexers, sample-and-hold circuits, and A/D converters.

When direct measurements are made, ultimate care must be exercised for the selection of the correct accelerometer to meet the requirements of a particular application. In order to reduce the errors, once the characteristics of the motion are studied, the following particulars of the accelerometers need to be considered: the transient response or cross-axis sensitivity; frequency range; sensitivity, mass and dynamic range; cross-axis response; and environmental conditions, such as temperature, cable noise, stability of bias, scale factor, and misalignment, etc.

Sensitivity of Accelerometers

During measurements, the transverse motions of the system affect most accelerometers and the sensitivity to these motions plays a major role in the accuracy of the measurement. The transverse, also known as cross-axis, sensitivity of an accelerometer is its response to acceleration in a plane perpendicular to the

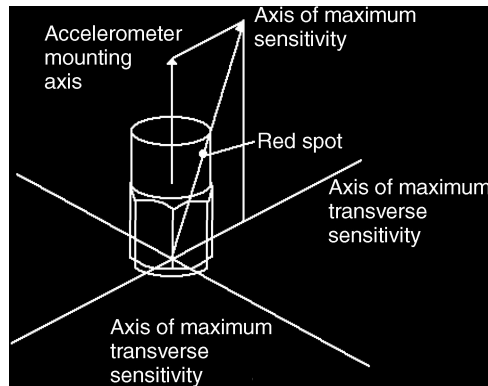



FIGURE 19.20 Illustration of cross-axis sensitivity.

main accelerometer axis as shown in Fig. 19.20. The cross-axis sensitivity is normally expressed as a percentage of the main-axis sensitivity and should be as low as possible. The manufacturers usually supply the direction of minimum sensitivity.

The measurement of the maximum cross-axis sensitivity is part of the individual calibration process and should always be less than 3% or 4%. If high levels of transverse vibration are present, this may result in erroneous overall results. In this case, separate arrangements should be made to establish the level and frequency contents of the cross-axis motions. The cross-axis sensitivity of typical accelerometers mentioned in the relevant sections were 2–3% for piezoelectric types and less than 1% for most others.

The Frequency Range

Acceleration measurements are normally confined to using the linear portion of the response curve of the accelerometer. The response is limited at the low frequencies as well as at the high frequencies by the natural resonances. As a rule of thumb, the upper-frequency limit for the measurement can be set to one-third of the accelerometer's resonance frequency such that the vibrations measured will be less than 1 dB in linearity. It should be noted that an accelerometer's useful frequency range may be significantly higher; that is, one-half or two-thirds of its resonant frequency. The measurement frequencies may be set to higher values in applications in which lower linearity (say 3 dB) is acceptable. The lower measuring frequency limit is determined by two factors. The first is the low-frequency cutoff of the associated preamplifiers. The second is the effect of ambient temperature fluctuations to which the accelerometer may be sensitive.

The Mass of Accelerometer and Dynamic Range

Ideally, the higher the transducer sensitivity, the better. Compromises may have to be made for sensitivity versus frequency, range, overload capacity, size, and so on.

In some cases, high errors will be introduced due to wrong selection of the sensor that is suitable for a specific application. For example, accelerometer mass becomes important when using small and light test objects. The accelerometer should not load the structural member, since additional mass can significantly change the levels and frequency presence at measuring points and invalidate the results. As a general rule, the accelerometer mass should not be greater than one-tenth the effective mass of the part or the structure that it is mounted onto for measurements.

The dynamic range of the accelerometer should match the high or low acceleration levels of the measured objects. General-purpose accelerometers can be linear from 5000*g* to 10,000*g*, which is well in the range of most mechanical shocks. Special accelerometers can measure up to 100,000*g*.

The Transient Response

Shocks are characterized as sudden releases of energy in the form of short-duration pulses exhibiting various shapes and rise times. They have high magnitudes and wide frequency contents. In applications where transient and shock measurements are involved, the overall linearity of the measuring system may be limited to high and low frequencies by a phenomena known as *zero shift* and *ringing*, respectively. The zero shift is caused by both the phase nonlinearity in the preamplifiers and the accelerometer not returning to steady-state operation conditions after being subjected to high shocks. Ringing is caused by high-frequency components of the excitation near-resonance frequency, preventing the accelerometer from returning back to its steady-state operation condition. To avoid measuring errors due to these effects, the operational frequency of the measuring system should be limited to the linear range.

Full Scale Range and Overload Capability

Most accelerometers are able to measure acceleration in both positive and negative directions. They are also designed to be able to accommodate overload capacity. Manufacturers also supply information on these two characteristics.

Environmental Conditions

In selection and implementation of accelerometers, environmental conditions such as temperature ranges, temperature transients, cable noise, magnetic field effects, humidity, and acoustic noise need to be considered. Manufacturers supply information on environmental conditions.

Inertial Accelerometers

Inertial accelerometers are mechanical accelerometers that make use of a seismic mass that is suspended by a spring or a lever inside a rigid frame as shown in Fig. 19.17. The frame carrying the seismic mass is connected firmly to the vibrating source whose characteristics are to be measured. As the system vibrates, the mass tends to remain fixed in its position so that the motion can be registered as a relative displacement between the mass and the frame. An appropriate transducer senses this displacement and the output signal is processed further. The displacement sensing element can be made from a variety of materials exhibiting resistive, capacitive, inductive, piezoelectric, piezoresistive, and optical capabilities. In practice, the seismic mass does not remain absolutely steady, but it can satisfactorily act as a reference position for selected frequencies.

By proper selection of mass, spring, and damper combinations, the seismic instrument may be used for either acceleration or displacement measurements. In general, a large mass and soft spring are suitable for vibration and displacement measurement, while a relatively small mass and a stiff spring are used in accelerometers. However, the term seismic is commonly applied to instruments, which sense very low levels of vibration in the ground or structures.

In order to describe the response of the seismic accelerometer in Fig. 19.17, Newton's second law, the equation of motion, may be written as

$$m\frac{d^2x_2}{dt^2} + c\frac{dx_2}{dt} + kx_2 = c\frac{dx_1}{dt} + kx_1 + mg\cos(\theta) \qquad (19.12)$$

where x_1 is the displacement of the vibration frame, x_2 is the displacement of the seismic mass, c is the velocity constant, θ is the angle between the sense axis and gravity, and k is the spring constant.

Taking md^2x_1/dt^2 from both sides of the equation and rearranging gives

$$m\frac{d^2z}{dt^2} + c\frac{dz}{dt} + kz = mg\cos(\theta) - m\frac{d^2x_1}{dt^2} \qquad (19.13)$$

where $z = x_2 - x_1$ is the relative motion between the mass and the base.

In Eq. (19.12), it is assumed that the damping force on the seismic mass is proportional to velocity only. If a harmonic vibratory motion is impressed on the instrument such that

$$x_1(t) = X_0 \sin(\omega_1 t) \tag{19.14}$$

where ω_1 is the frequency of vibration (rad/s), writing

$$-m\frac{d^2x_1}{dt^2} = mX_0 \sin\omega_1 t$$

modifies Eq. (19.13) as

$$m\frac{d^2z}{dt^2} + c\frac{dz}{dt} + kz = mg\cos(\theta) + ma_1\sin\omega_1 t \tag{19.15}$$

where $a_1 = mX_0\,\omega_1^2$.

Equation (19.15) will have transient and steady-state solutions. The steady-state solution of the differential equation (19.15) may be determined as

$$z = \frac{mg\cos(\theta)}{k} + \frac{ma_1\sin\omega_1 t}{(k - m\omega_1^2 + jc\omega_1)} \tag{19.16}$$

Rearranging Eq. (19.16) results in

$$z = \frac{mg\cos(\theta)}{\omega_n} + \frac{a_1\sin(\omega_1 t - \phi)}{\sqrt{\omega_n^2(1-r^2)^2 + (2\zeta r)^2}} \tag{19.17}$$

where ω_n (=$\sqrt{k/m}$) is the natural frequency of the seismic mass, ζ (=$c/2\sqrt{km}$) is the damping ratio. The damping ratio can be written in terms of the critical damping ratio as $\zeta = c/c_c$, where $c_c = 2\sqrt{km}$, ϕ (=$\tan^{-1}[c\omega_1/(k - m\omega_1^2)]$) is the phase angle, and r (=ω_1/ω_n) is the frequency ratio.

A plot of Eq. (19.17), $(x_1 - x_2)_0/x_0$ against frequency ratio ω_1/ω_n, is illustrated in Fig. 19.21. This figure shows that the output amplitude is equal to the input amplitude when $c/c_c = 0.7$ and $\omega_1/\omega_n > 2$. The output becomes essentially a linear function of the input at high frequency ratios. For satisfactory system performance, the instrument constant c/c_c and ω_n should be carefully calculated or obtained from calibrations.

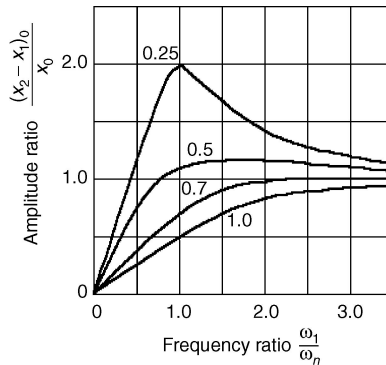



FIGURE 19.21 Frequency versus amplitude ratio of seismic accelerometers.

In this way the anticipated accuracy of measurement may be predicted for frequencies of interest. A comprehensive treatment of the analysis has been performed by McConnell [7]; interested readers should refer to this text for further details.

Seismic instruments are constructed in a variety of ways. In a potentiometric instrument, a voltage divider potentiometer is used for sensing the relative displacement between the frame and the seismic mass. In the majority of potentiometric accelerometers, the device is filled with a viscous liquid that interacts continuously with the frame and the seismic mass to provide damping. These accelerometers have a low frequency of operation (less than 100 Hz) and are mainly intended for slowly varying accelerations, and low-frequency vibrations. A typical family of such instruments offers many different models, covering the range of ±1*g* to ±50*g* full scale. The natural frequency ranges from 12 to 89 Hz, and the damping ratio ζ can be kept between 0.5 and 0.8 by using a temperature compensated liquid-damping arrangement. Potentiometer resistance may be selected in the range of 1000–10,000 W, with a corresponding resolution of 0.45–0.25% of full scale. The cross-axis sensitivity is less than ±1%. The overall accuracy is ±1% of full scale or less at room temperatures. The size is about 50 mm^3 with a total mass of about 1/2 kg.

Linear variable differential transformers (LVDTs) offer another convenient means of measurement of the relative displacement between the seismic mass and the accelerometer housing. These devices have higher natural frequencies than potentiometer devices, up to 300 Hz. Since the LVDT has lower resistance to motion, it offers much better resolution. A typical family of liquid-damped differential-transformer accelerometers exhibits the following characteristics. The full scale ranges from ±2*g* to ±700*g*, the natural frequency from 35 to 620 Hz, the nonlinearity 1% of full scale. The full-scale output is about 1 V with an LVDT excitation of 10 V at 2000 Hz, the damping ratio ranges from 0.6 to 0.7, the residual voltage at the null position is less than 1%, and the hysteresis is less than 1% of full scale. The size is about 50 mm^3 with a mass of about 120 g.

Electrical resistance strain gages are also used for displacement sensing of the seismic mass. In this case, the seismic mass is mounted on a cantilever beam rather than on springs. Resistance strain gauges are bonded on each side of the beam to sense the strain in the beam resulting from the vibrational displacement of the mass. A viscous liquid that entirely fills the housing provides damping of the system. The output of the strain gauges is connected to an appropriate bridge circuit. The natural frequency of such a system is about 300 Hz. The low natural frequency is due to the need for a sufficiently large cantilever beam to accommodate the mounting of the strain gauges.

One serious drawback of the seismic instruments is temperature effects requiring additional compensation circuits. The damping of the instrument may also be affected by changes in the viscosity of the fluid due to temperature. For instance, the viscosity of silicone oil, often used in these instruments, is strongly dependent on temperature.

Suspended-Mass, Cantilever, and Pendulum-Type Inertial Accelerometers

There are a number of different inertial-type accelerometers, most of which are in development stages or used under very special circumstances, such as gyropendulum, reaction-rotor, vibrating-string, and centrifugal-force-balance.

The vibrating-string instrument, Fig. 19.22, makes use of a proof mass supported longitudinally by a pair of tensioned, transversely vibrating strings with uniform cross section and equal lengths and masses. The frequency of vibration of the strings is set to several thousand cycles per second. The proof mass is supported radially in such a way that the acceleration normal to the strings does not affect the string tension. In the presence of acceleration along the sensing axis, a deferential tension exists on the two strings, thus altering the frequency of vibration. From the second law of motion the frequencies may be written as

$$f_1^2 = \frac{T_1}{4m_s l} \qquad \text{and} \qquad f_2^2 = \frac{T_2}{4m_s l} \tag{19.18}$$

where T is the tension, m_s are the mass, and l is the lengths of strings.

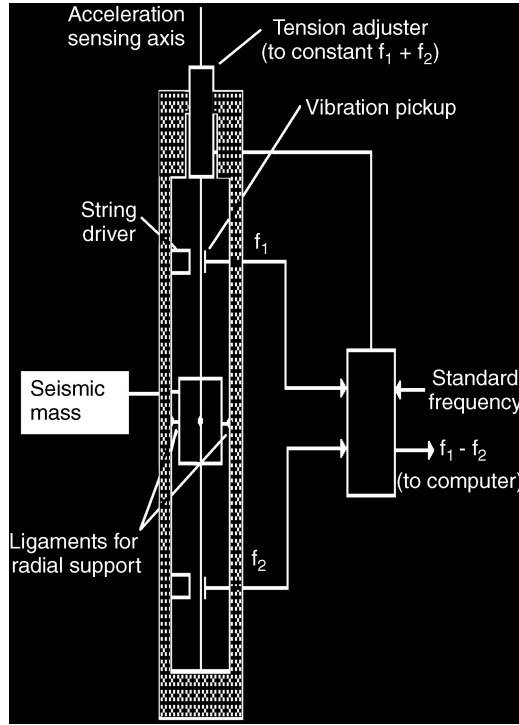



FIGURE 19.22 A typical suspended-mass-vibrating string accelerometer.

The quantity $T_1 - T_2$ is proportional to *ma* where *a* is the acceleration along the axis of the strings. An expression for the difference of the frequency-squared terms may be written as

$$f_1^2 - f_2^2 = \frac{T_1 - T_2}{4m_s l} = \frac{ma}{4m_s l} \tag{19.19}$$

Hence

$$f_1 - f_2 = \frac{ma}{4m_s l(f_1 + f_2)} \tag{19.20}$$

The sum of frequencies $(f_1 + f_2)$ can be held constant by serving the tension in the strings with reference to the frequency of a standard oscillator. Then, the difference between the frequencies becomes linearly proportional to acceleration. In some versions, the beamlike property of the vibratory elements is used by gripping them at nodal points corresponding to the fundamental mode of the vibration of the beam, and at the respective centers of percussion of the common proof mass. The output frequency is proportional to acceleration and the velocity is proportional to phase, thus offering an important advantage. The improved versions of these devices lead to cantilever-type accelerometers, discussed next.

In a cantilever-type accelerometer, a small cantilever beam mounted on the block is placed against the vibrating surface, and an appropriate mechanism is provided for varying the beam length. The beam length is adjusted such that its natural frequency is equal to the frequency of the vibrating surface, and hence the resonance condition is obtained. Recently, slight variations of cantilever-beam arrangements are finding new applications in microaccelerometers.

In a different type of suspended-mass configuration, a pendulum is used that is pivoted to a shaft rotating about a vertical axis. Pickoff mechanisms are provided for the pendulum and the shaft speed.

The system is servo controlled to maintain it at null position. Gravitational acceleration is balanced by the centrifugal acceleration. The shaft speed is proportional to the square root of the local value of the acceleration.

Electromechanical Accelerometers

Electromechanical accelerometers, essentially servo or null-balance types, rely on the principle of feedback. In these instruments, an acceleration-sensitive mass is kept very close to a neutral position or zero displacement point by sensing the displacement and feeding back the effect of this displacement. A proportional magnetic force is generated to oppose the motion of the mass displaced from the neutral position, thus restoring this position just as a mechanical spring in a conventional accelerometer would do. The advantages of this approach are better linearity and elimination of hysteresis effects, as compared to the mechanical springs. Also, in some cases, electrical damping can be provided, which is much less sensitive to temperature variations.

One very important feature of electromechanical accelerometers is the capability of testing the static and dynamic performances of the devices by introducing electrically excited test forces into the system. This remote self-checking feature can be quite convenient in complex and expensive tests where accuracy is essential. These instruments are also useful in acceleration control systems, since the reference value of acceleration can be introduced by means of a proportional current from an external source. They are used for general-purpose motion measurements and monitoring low-frequency vibrations.

There are a number of different electromechanical accelerometers: coil-and-magnetic types, induction types, etc.

Coil-and-Magnetic Accelerometers

These accelerometers are based on Ampere's law, that is, "a current-carrying conductor disposed within a magnetic field experiences a force proportional to the current, the length of the conductor within the field, the magnetic field density, and the sine of the angle between the conductor and the field." The coils of these accelerometers are located within the cylindrical gap defined by a permanent magnet and a cylindrical soft iron flux return path. They are mounted by means of an arm situated on a minimum friction bearing or flexure so as to constitute an acceleration-sensitive seismic mass. A pickoff mechanism senses the displacement of the coil under acceleration and causes the coil to be supplied with a direct current via a suitable servo controller to restore or maintain a null condition. The electrical currents in the restoring circuit are linearly proportional to acceleration, provided (1) armature reaction affects are negligible and fully neutralized by a compensating coil in opposition to the moving coil, and (2) the gain of the servo system is large enough to prevent displacement of the coil from the region in which the magnetic field is constant.

In these accelerometers, the magnetic structure must be shielded adequately to make the system insensitive to external disturbances or the earth's magnetic field. Also, in the presence of acceleration there will be a temperature rise due to i^2R losses. The effects of these i^2R losses on the performance are determined by the thermal design and heat-transfer properties of the accelerometers.

Induction Accelerometers

The cross-product relationship of current, magnetic field, and force is the basis for induction-type electromagnetic accelerometers. These accelerometers are essentially generators rather than motors. One type of instrument, the cup-and-magnet design, includes a pendulous element with a pickoff mechanism and a servo controller driving a tachometer coupling. A permanent magnet and a flux return ring, closely spaced with respect to an electrically conductive cylinder, are attached to the pendulous element. A rate-proportional drag force is obtained by the electromagnetic induction effect between the magnet and the conductor. The pickoff mechanism senses pendulum deflection under acceleration and causes the servo controller to turn the rotor to drag the pendulous element toward the null position. Under steady-state conditions motor speed is a measure of the acceleration acting on the instrument. Stable servo operation is achieved by employing a time-lead network to compensate the inertial time lag of

the motor and magnet combination. The accuracy of the servo-type accelerometers is ultimately limited by consistency and stability of scale factors of coupling and cup-and-magnet devices as a function of time and temperature.

Another accelerometer based on induction design uses the eddy-current induction torque generation. The force-generating mechanism of an induction accelerometer consists of a stable magnetic field, usually supplied by a permanent magnet, which penetrates orthogonally through a uniform conduction sheet. The movement of the conducting sheet relative to the magnetic field in response to acceleration results in a generated electromotive potential in each circuit in the conductor. This action is in accordance with Faraday's principle. In induction-type accelerometers, the induced eddy currents are confined to the conductor sheet, making the system essentially a drag coupling. Since angular rate is proportional to acceleration, angular position represents change in velocity. This is a particularly useful feature in navigation applications.

A typical commercial instrument based on the servo-accelerometer principle might have a micromachined quartz flexure suspension, differential capacitance angle pick-off, air-squeeze film plus servo-lead compensation for system damping. Of the available models, as an example, a typical 30*g* unit has a threshold and resolution of 1 μg, a frequency response that is flat to within 0.05% at 10 Hz and 2% at 100 Hz, a natural frequency of 1500 Hz, a damping ratio from 0.3 to 0.8, and transverse or cross-axis sensitivity of 0.1%. If, for example, the output current is about 1.3 mA/*g*, a 250 W readout resistor would give about ±10 V full scale at 30*g*. These accelerometers are good for precision work and used in many applications such as aircraft and missile control systems, measurement of tilt angles for borehole navigation, and axle angular bending in aircraft weight and balance systems.

Piezoelectric Accelerometers

Piezoelectric accelerometers are widely used for general-purpose acceleration, shock, and vibration measurements. They are basically motion transducers with large output signals and comparatively small sizes and they are self generators not requiring external power sources. They are available with very high natural frequencies and are therefore suitable for high-frequency applications and shock measurements.

These devices utilize a mass in direct contact with the piezoelectric component or crystal as shown in Fig. 19.23. When a varying motion is applied to the accelerometer, the crystal experiences a varying force excitation ($F = ma$), causing a proportional electric charge q to be developed across it. So,

$$q = d_{ij}F = d_{ij}ma \tag{19.21}$$

where q is the charge developed and d_{ij} is the piezoelectric coefficient of the material.

As this equation shows, the output from the piezoelectric material is dependent on its mechanical properties, d_{ij}. Two commonly used piezoelectric crystals are lead-zirconate titanate ceramic (PZT) and crystalline quartz. They are both self-generating materials and produce a large electric charge for their size. The piezoelectric strain constant of PZT is about 150 times that of quartz. As a result, PZTs are much more sensitive and smaller in size than quartz counterparts. These accelerometers are useful for

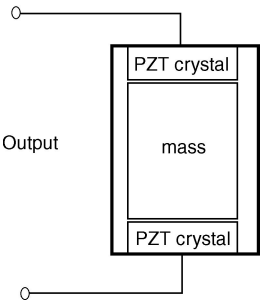



FIGURE 19.23 A compression type piezoelectric accelerometer arrangement.

high-frequency applications. The roll-off typically starts near 100 Hz. These active devices have no DC response. Since piezoelectric accelerometers have comparatively low mechanical impedances, their effect on the motion of most structures is negligible.

Mathematically, their transfer function approximates a third-order system that can be expressed as

$$\frac{e_0(s)}{a(s)} = \frac{K_q \tau s}{C\omega_n^2(\tau s + 1)(s^2/\omega_n^2 + 2\zeta s/\omega_n + 1)} \tag{19.22}$$

where K_q is the piezoelectric constant related to charge (C cm), τ is the time constant of the crystal, and s is the Laplace variable. It is worth noting that the crystal itself does not have a time constant τ, but the time constant is observed when the accelerometer is connected to an electric circuit, for example, an *RC* circuit.

The low-frequency response is limited by the piezoelectric characteristic $\tau s/(\tau s + 1)$, while the high-frequency response is related to mechanical response. The damping factor ζ is very small, usually less than 0.01 or near zero. Accurate low-frequency response requires large τ, which is usually achieved by use of high-impedance voltage amplifiers. At very low frequencies thermal effects can have severe influences on the operation characteristics.

In piezoelectric accelerometers, two basic design configurations are used: compression types and shear-stress types. In compression-type accelerometers, the crystal is held in compression by a preload element; therefore the vibration varies the stress in compressed mode. In a shear-stress accelerometer, vibration simply deforms the crystal in shear mode. The compression accelerometer has a relatively good mass to sensitivity ratio and hence exhibits better performance. But, since the housing acts as an integral part of the spring-mass system, it may produce spurious interfaces in the accelerometer output if excited around its natural frequency.

Piezoelectric accelerometers are available in a wide range of specifications and are offered by a large number of manufacturers. For example, the specifications of a shock accelerometer may have 0.004 pC/*g* in sensitivity and a natural frequency of up to 250,000 Hz, while a unit designed for low-level seismic measurements might have 1000 pC/*g* in sensitivity and only 7000 Hz natural frequency. They are manufactured as small as 3×3 mm in dimension with about 0.5 g in mass, including cables. They have excellent temperature ranges and some of them are designed to survive the intensive radiation environment of nuclear reactors. However, piezoelectric accelerometers tend to have larger cross-axis sensitivity than other types, about 2–4%. In some cases, large cross-axis sensitivity may be minimized during installations by the correct orientation of the device. These accelerometers may be mounted with threaded studs, with cement or wax adhesives, or with magnetic holders.

Piezoresistive Accelerometers

Piezoresistive accelerometers are essentially semiconductor strain gauges with large gauge factors. High gauge factors are obtained since the material resistivity is dependent primarily on the stress, not only on the dimensions. This effect can be greatly enhanced by appropriate doping of semiconductors such as silicon. Most piezoresistive accelerometers use two or four active gauges arranged in a Wheatstone bridge. Extra precision resistors are used, as part of the circuit, in series with the input to control the sensitivity, for balancing, and for offsetting temperature effects. The sensitivity of a piezoresistive sensor comes from the elastic response of its structure and resistivity of the material. Wire and thick or thin film resistors have low gauge factors, that is, the resistance change due to strain is small. The mechanical construction of a piezoresistive accelerometer is shown in Fig. 19.24.

Piezoresistive accelerometers are useful for acquiring vibration information at low frequencies, for example, below 1 Hz. In fact, they are inherently true non-vibrational acceleration sensors. They generally have wider bandwidth, smaller nonlinearities and zero shifting, and better hysteresis characteristics compared to piezoelectric counterparts. They are suitable to measure shocks well above 100,000*g*. Typical characteristics

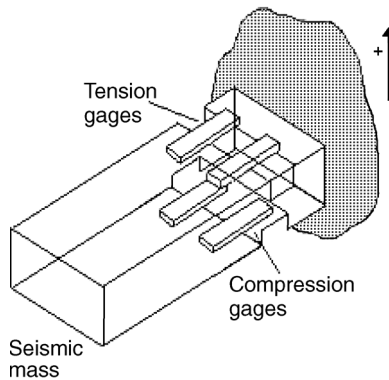



FIGURE 19.24 Bonding of piezoresistive and piezoelectric accelerometers to the inertial systems.

of piezoresistive accelerometers may be listed to be 100 mV/*g* as the sensitivity, 0–750 Hz as the frequency range, 2500 Hz in resonance frequency, 25*g* as the amplitude range, 2000*g* as the shock rating, and 0–95°C as the temperature range, with a total mass of about 25 g.

Most contemporary piezoresistive sensors are manufactured from a single piece of silicon. This gives better stability and less thermal mismatch between parts. In a typical monolithic sensing element a 1-mm silicon chip incorporates the spring, mass, and four-arm bridge assembly. The elements are formed by a pattern of dopant in the originally flat silicon. Subsequent etching of channels frees the gauges and simultaneously defines the masses as regions of silicon of original thickness.

Strain-Gauge Accelerometers

Strain-gauge accelerometers are based on resistance properties of electrical conductors. If a conductor is stretched or compressed, its resistance alters due to (a) dimensional changes, and (b) the changes in the fundamental property of material called piezoresistance. This indicates that the resistivity ρ of the conductor depends on the mechanical strain applied onto it. The dependence is expressed as the gauge factor:

$$\frac{dR/R}{dL/L} = 1 + 2\nu + \frac{d\rho/\rho}{dL/L} \tag{19.23}$$

where 1 indicates the resistance change due to length, 2ν indicates resistance change due to area, and $(d\rho/\rho)/(dL/L)$ indicates the resistance change due to piezoresistivity.

There are many types of strain-gauges: unbonded metal-wire gauges, bonded metal-wire gauges, bonded metal-foil gauges, vacuum-deposited thin-metal-film gauges, bonded semiconductor gauges, and diffused semiconductor gauges. However, usually bonded and unbonded metal-wire gauges find wider applications. A section of the strain-gauge accelerometers, particularly bonded semiconductor types, known as the piezoresistive transducers, are used, but they suffer from high temperature sensitivities, nonlinearities, and some mounting difficulties. Nevertheless, with the recent developments of micromachine technology, these sensors have been improved considerably, thus finding many new applications.

Unbonded-strain-gauge accelerometers use the strain wires as the spring element and as the motion transducer, using similar arrangements as in Fig. 19.25. They are useful for general-purpose motion and vibration measurements from low to medium frequencies. They are available in wide ranges and characteristics: typically ±5*g* to ±200*g* full scale, a natural frequency of 17–800 Hz, a 10-V excitation voltage AC or DC, full scale output ±20 mV to ±50 mV, a resolution less than 0.1%, an inaccuracy less than 1% full scale, and a cross-axis sensitivity less than 2%. The damping ratio (using silicone oil damping) is 0.6–0.8 at room temperature. These instruments are small and light, usually with a mass less than 25 g.

Bonded-strain-gauge accelerometers generally use a mass supported by a thin flexure beam. The strain gauges are cemented onto the beam to achieve maximum sensitivity, temperature compensation, and sensitivity to both cross-axis and angular accelerations. Their characteristics are similar to the unbonded-strain-gauge accelerometers but have greater sizes and weights. Often silicone oil is used for damping. Semiconductor strain gauges are widely used as strain sensors in cantilever-beams and mass types of accelerometers. They allow high outputs (0.2–0.5 V full scale). Typically, a ±25*g* acceleration unit has a flat response from 0 to 750 Hz, a damping ratio of 0.7, a mass of about 28 g, and an operational temperature of −18°C to +93°C. A triaxial ±20,000*g* model has a flat response from 0 to 15 kHz, a damping ratio of 0.01, and a compensation temperature range of 0–45°C, and is $13 \times 10 \times 13$ mm^3 in size and 10 g in mass.

Electrostatic Accelerometers

Electrostatic accelerometers are based on Coulomb's law between two charged electrodes; therefore, they are capacitive types. Depending on the operation principles and external circuits they can be broadly classified as (a) electrostatic-force-feedback accelerometers, and (b) differential-capacitance accelerometers.

Electrostatic-Force-Feedback Accelerometers

An electrostatic-force-feedback accelerometer consists of an electrode, with mass m and area S, mounted on a light pivoted arm that moves relative to some fixed electrodes. The nominal gap h between the pivoted and fixed electrodes is maintained by means of a force-balancing servo system, which is capable of varying the electrode potential in response to signals from a pickoff mechanism that senses relative changes in the gap. Mathematically, the field between the electrodes may be expressed by

$$E = \frac{Q}{\varepsilon k S} \tag{19.24}$$

where E is the intensity or potential gradient (dV/dx), Q is the charge, S is the area of the conductor, and k is the dielectric constant of the space outside the conductor.

From this expression, it can be shown that the force per unit area of the charged conductor (in N/m^2) is given by

$$\frac{F}{S} = \frac{Q^2}{2\varepsilon k S^2} = \frac{\varepsilon k E^2}{2} \tag{19.25}$$

Consider one movable and one stationary electrode and assume that the movable electrode is maintained at a bias potential V_1 and the stationary one at a potential V_2. The electrical intensity E in the gap, h, can be expressed as

$$E_1 = \frac{V_1 - V_2}{h} \tag{19.26}$$

so that the force of attraction may be found as

$$F_1 = \frac{\varepsilon k E^2 S}{2h^2} = \frac{\varepsilon k (V_1 - V_2)^2 S}{2h^2} \tag{19.27}$$

In the presence of acceleration, if V_2 is adjusted to restrain the movable electrode to the null position, the expression relating acceleration and electrical potential may be given by

$$a = \frac{F_1}{m} = \frac{\varepsilon k (V_1 - V_2)^2 S}{2h^2 m} \tag{19.28}$$

The device so far described can measure acceleration in one direction only, and the output is quadratic in character, that is,

$$(V_1 - V_2) = D\sqrt{a} \tag{19.29}$$

where D is the constant of proportionality.

The output may be linearized in a number of ways, one of which is the quarter-square method. If the servo controller applies a potential $-V_2$ to the other fixed electrode, the force of attraction between this electrode and the movable electrode becomes

$$a = \frac{F_2}{m} = \frac{\varepsilon k(V_1 + V_2)^2 S}{2h^2 m} \tag{19.30}$$

and the force-balance equation of the movable electrode when the instrument experiences a downward acceleration a now is

$$ma = F_2 - F_1 = \frac{\varepsilon kS[(V_1 + V_2)^2 - (V_1 - V_2)^2]}{2h^2}$$

or

$$ma = F_2 - F_1 = \frac{2\varepsilon kSV_1V_2}{h^2} \tag{19.31}$$

Hence, if the bias potential V_1 is held constant and the gain of the control loop is high so that variations in the gap are negligible, the acceleration becomes a linear function of the controller output voltage V_2.

The principal difficulty in mechanizing the electrostatic force accelerometer is the relatively high electric field intensity required to obtain an adequate force. Damping can be provided electrically or by viscosity of the gaseous atmosphere in the inter-electrode space if the gap h is sufficiently small. The scheme works best in micromachined instruments. Nonlinearity in the voltage breakdown phenomenon permits larger gradients in very small gaps.

A typical electrostatic accelerometer has the following characteristics: range ±50*g*, resolution $10^{-3}g$, sensitivity 100 mV/*g*, nonlinearity <1% FS, transverse sensitivity <1% FS, thermal sensitivity 6×10^{-4}/K, mechanical shock 10,000*g*, operating temperature –45°C to 90°C, supply voltage 5 V DC, and weight 45 g. The main advantages of electrostatic accelerometers are their extreme mechanical simplicity, low power requirements, absence of inherent sources of hysteresis errors, zero temperature coefficients, and ease of shielding from stray fields.

Differential-Capacitance Accelerometers

Differential-capacitance accelerometers are based on the principle of the change of capacitance in proportion to applied acceleration. In one type, the seismic mass of the accelerometer is made as the movable element of an electrical oscillator. The seismic mass is supported by a resilient parallel-motion beam arrangement from the base. The system is set to have a certain defined nominal frequency when undisturbed. If the instrument is accelerated, the frequency varies above and below the nominal value depending on the direction of acceleration.

The seismic mass carries an electrode located in opposition to a number of base-fixed electrodes that define variable capacitors. The base-fixed electrodes are resistances coupled in the feedback path of a wideband, phase-inverting amplifier. The gain of the amplifier is predetermined to ensure maintenance

of oscillations over the range of variation of the capacitance determined by the applied acceleration. The value of the capacitance C for each of the variable capacitors is given by

$$C = \frac{\varepsilon k S}{h} \tag{19.32}$$

where k is the dielectric constant, ε is the permittivity of free space, S is the area of the electrode, and h is the variable gap.

Denoting the magnitude of the gap for zero acceleration as h_0, the value of h in the presence of acceleration a may be written as

$$h = h_0 + \frac{ma}{K} \tag{19.33}$$

where m is the value of the proof mass and K is the spring constant. Thus,

$$C = \frac{\varepsilon k S}{h_0 + (ma/K)} \tag{19.34}$$

For example, the frequency of oscillation of the resistance-capacitance type circuit is given by the expression

$$f = \frac{\sqrt{6}}{2\pi RC} \tag{19.35}$$

Substituting this value of C in Eq. (19.34) gives

$$f = \frac{\sqrt{6}[h_0 + (ma/K)]}{2\pi R \varepsilon k S} \tag{19.36}$$

Denote the constant quantity $\sqrt{6}/(2\pi R \varepsilon k S)$ as B and rewrite Eq. (19.36) as

$$f = Bh_0 + \frac{Bma}{K} \tag{19.37}$$

The first term on the right-hand side expresses the fixed bias frequency f_0 and the second term denotes the change in frequency resulting from acceleration, so that the expression may be written as

$$f = f_0 + f_a \tag{19.38}$$

If the output frequency is compared with an independent source of a constant frequency of f_0, then f_a can be determined easily.

A commonly used capacitive-type accelerometer is based on a thin diaphragm with spiral flexures that provide the spring, proof mass, and moving plate necessary for the differential capacitor. Plate motion between the electrodes pumps air parallel to the plate surface and through holes in the plate to provide squeeze film damping. Since air viscosity is less temperature sensitive than oil, the desired damping ratio of 0.7 hardly changes more than 15%. A family of such instruments are easily available with full-scale ranges from ±0.2g (4 Hz flat response) to ±1000g (3000 Hz), a cross-axis sensitivity less than 1%, and a full-scale output of ±1.5 V. The size of a typical device is about 25 mm^3 with a mass of 50 g.

Micro- and Nanoaccelerometers

By the end of the 1970s it became apparent that the essentially planar processing integrated-circuit (IC) technology could be modified to fabricate three-dimensional electromechanical structures by the micromachining process. Accelerometers and pressure sensors were among the first IC sensors. The first accelerometer was developed in 1979. Since then the technology has been progressing steadily and now an extremely diverse range of accelerometers are readily available. Most sensors use bulk micromachining rather than surface micromachining techniques. In bulk micromachining the flexures, resonant beams, and all other critical components of the accelerometer are made from bulk silicon in order to exploit the full mechanical properties of silicon crystals. With proper design and film process, bulk micromachining yields extremely stable and robust accelerometers.

The selective etching of multiple layers of deposited thin films, or surface micromachining, allows movable microstructures to be fabricated on silicon wafers. With surface micromachining, layers of structure material are disposed and patterned as shown in Fig. 19.25. These structures are formed by polysilicons and sacrificial materials such as silicon dioxides. The sacrificial material acts as an intermediate spacer layer and is etched away to produce a freestanding structure. Surface machining technology also allows smaller and more complex structures to be built in multiple layers on a single substrate. A typical example of modern micromachined accelerometer is given in Fig. 19.26. Multiple accelerometers can be mounted on a single chip, sensing accelerations in x, y, and z directions. The primary signal conditioning is also provided in the same chip. The output from the chip is usually read in the digital form.

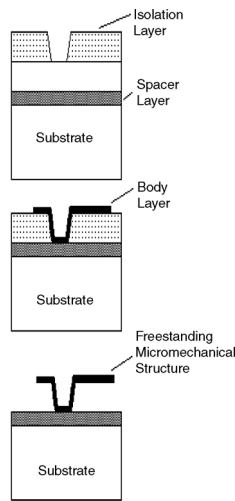



FIGURE 19.25 Steps of micromachining to manufacture micro- and nanoaccelerometers.

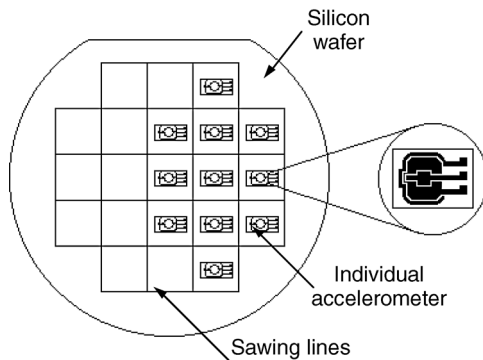



FIGURE 19.26 Multiple accelerometers in a single chip.

Most micro- and nanoaccelerometers detect acceleration by measuring the relative motion between proof mass and mounting substrate. The proof mass is suspended above the substrate by a mechanical spring suspension. When the sensor undergoes acceleration, the proof mass tends to remain stationary and therefore displaces with respect to the moving substrate. This displacement is measured capacitively or by means of piezoresistive or piezoelectric methods using CMOS technology. Chip circuits provide offset cancellation for bias stability, gain scale factor stability, zero acceleration bias stability, temperature compensation, prefiltering, noise immune digital output, and so on.

The operational principles of some of the microaccelerometers are very similar to capacitive force-balance or vibrating-beam accelerometers, discussed earlier. Manufacturing techniques may change from one manufacturer to another. However, in general, vibrating-beam accelerometers are preferred because of better air-gap properties and improved bias performance characteristics.

Vibrating-beam accelerometers, also termed resonant-beam force transducers, are made in such a way that an acceleration along a positive input axis places the vibrating beam in tension. Thus, the resonant frequency of the vibrating beam increases or decreases with the applied acceleration.

In DETF, an electronic oscillator capacitively couples energy into two vibrating beams to keep them oscillating at their resonant frequency. The beams vibrate 180° out of phase to cancel reaction forces at the ends. The dynamic cancellation effect of the DETF design prevents energy from being lost through the ends of the beam. Hence, the dynamically balanced DETF resonator has a high Q factor, which leads to a stable oscillator circuit. The acceleration signal is produced from the oscillator as a frequency-modulated square wave that can be used for a digital interface.

The frequency of resonance of the system must be much higher than any input acceleration, and this limits the measurable range. In a micromachined accelerometer, used in military applications, the following characteristics are given: a range of $\pm 1200g$, a sensitivity of 1.11 Hz/g, a bandwidth of 2500 Hz, an unloaded DETF frequency of 9952 Hz. The frequency at $+1200g$ is 11,221 Hz, the frequency at $-1200g$ is 8544 Hz, and the temperature sensitivity is 5 mg/°C. The accelerometer size is 6 mm diameter by 4.3 mm length, with a mass of about 9 g. It has a turn-on time of less than 60 s, the accelerometer is powered with +9 to +16 V DC, and the nominal output is a 9000-Hz square wave.

Surface micromachining has also been used to manufacture specific application accelerometers, such as air-bag applications in the automotive industry. In one type, a three-layer differential capacitor is created by alternate layers of polysilicon and phosphosilicate glass (PSG) on a 0.38-mm thick, 100-mm long wafer. A silicon wafer serves as the substrate for the mechanical structure. The trampoline-shaped middle layer is suspended by four supporting arms. This movable structure is the seismic mass for the accelerometer. The upper and lower polysilicon layers are fixed plates for the differential capacitors. The glass is sacrificially etched by hydrofluoric acid (HF).

Signal Conditioning and Biasing

Common signal conditioners are appropriate for interfacing accelerometers to computers or other instruments for further signal processing. Generally, the generated raw signals are amplified and filtered suitably by the circuits within the accelerometer casing supplied by manufacturers. Nevertheless, piezoelectric and piezoresistive transducers require special signal conditioners with certain characteristics that will be discussed next.

Piezoelectric Accelerometers

Piezoelectric accelerometers supply small energy to the signal conditioners since they have high capacitive source impedances. The equivalent circuit of a piezoelectric accelerometer can be regarded as an active capacitor that charges itself when mechanically loaded. The selection of the elements of the external signal conditioning circuit is dependent on the characteristics of the equivalent circuit. A most common approach is the charge amplifier since the system gain and low-frequency responses of these amplifiers are well defined. The performance of the circuit is also independent of cable length and capacitance of the accelerometer. In many applications, noise-treated cables are necessary to avoid the triboelectric charges occurring due to movement of cables.

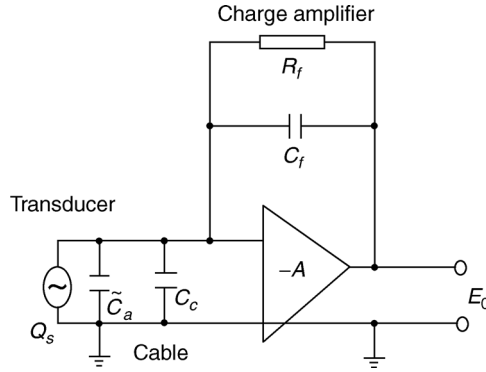



FIGURE 19.27 A typical charge amplifier.

The charge amplifier (with about 1000 MΩ input impedance) basically converts the input charge to voltage first and then amplifies this voltage. It consists of a charge converter output voltage, which occurs as a result of the charge input signal returning through the feedback capacitor to maintain the input voltage at the input level close to zero, as shown in Fig. 19.27. With the help of basic operational-type feedback, the amplifier input is maintained at essentially 0 V; therefore, it looks like a short circuit to the input. Thus, the charge input is stored in the feedback capacitor, producing a voltage across it that is equal to the value of the charge input divided by the capacitance of the feedback capacitor. The complete transfer function of the circuit describing the relationship between the output voltage and the input acceleration magnitude may be determined by the following complex transform:

$$\frac{E_0}{a_0} = S_a jR_f C_f \omega\left[1 + jR_f C_f \omega\left(1 + C_f + \frac{C_a + C_c}{1+G}\right)\right] \tag{19.39}$$

where E_0 is the charge converter output (V), a_0 is the magnitude of acceleration (m/s^2), S_a is the accelerometer sensitivity (mV/g), C_a is the accelerometer capacitance (F), C_c is the cable capacitance (F), C_f is the feedback capacitance (F), R_f is the feedback loop resistance, and G is the amplifier open-loop gain.

In most applications, since C_f is selected to be large compared to $(C_a + C_c)/(1 + G)$, the system gain becomes independent of the cable length. In this case the denominator of the equation can be simplified to give a first-order system with roll off at

$$f_{-3\text{dB}} = \frac{1}{2\pi R_e C_f} \tag{19.40}$$

with a slope of 10 dB per decade. For practical purposes, the low-frequency response of this system is a function of well-defined electronic components and does not vary by cable length. This is an important feature when measuring low-frequency vibrations.

Many piezoelectric accelerometers are manufactured with preamplifiers and other signal-conditioning circuits enclosed in the same casing. Some accelerometer preamplifiers include integrators to convert the acceleration proportional outputs to either velocity or displacement proportional signals. To attenuate noise and vibration signals that lie outside the frequency range of interest, most preamplifiers are equipped with a range of high-pass and low-pass filters. This avoids interference from electrical noise or signals inside the linear portion of the accelerometer frequency range. Nevertheless, it is worth mentioning that these devices usually have two time constants, external and internal. The mixture of these two time constants can lead to problems particularly at low frequencies. Manufacturers through design and construction usually fix the internal time constants. However, care must be observed to account for the effect of external time constants through impedance matching.

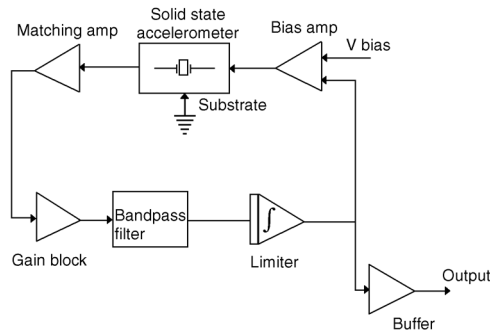



FIGURE 19.28 A typical signal conditioning arrangement for single chip microaccelerometers.

Piezoresistive Transducers

Piezoresistive transducers generally have high-amplitude outputs, low-output impedance, and low intrinsic noise. Most of these transducers are designed for constant-voltage excitations. They are usually calibrated for constant-current excitations to avoid external interference. Many piezoresistive transducers are configured as full-bridge devices. Some have four active piezoresistive arms, together with two fixed precision resistors to permit shunt calibration.

Microaccelerometers

In microaccelerometers signal-conditioning circuitry is integrated within the same chip as the sensor. A typical example of the signal-conditioning circuitry is given in Fig. 19.28 in block diagram form. In this type of accelerometer, the electronic system is essentially a crystal-controlled oscillator circuit and the output signal of the oscillator is a frequency-modulated acceleration signal. Some circuits provide a buffered square-wave output that can be directly interfaced digitally. In these cases the need for analog-to-digital (A/D) conversion is eliminated, thus removing one of the major sources of errors. In other types of accelerometers, signal conditioning circuits such as A/D converters are retained within the chip.

Force Feedback Accelerometers

Signals from force feedback accelerometers often must be digitized for use in digital systems. A common solution is to use voltage to frequency or current to frequency converters to convert the analog signals to train pulses. These converters are expensive, often as much as the accelerometer, and add as much to the error budget.

Here, it is worth mentioning that GPS systems are becoming add-ons to many position sensing mechanisms. Because of antenna dynamics, shadowing, multipath effects, and to provide redundancy for critical systems such as aircraft, many of these systems require inertial aiding, tied-in with accelerometers and gyros. With the development of micromachining, small and cost-effective GPS assisted inertial systems will be available in the near future. These developments will require extensive signal processing with a high degree of accuracy. Dynamic ranges on the order of a million to one (e.g., 30–32 bits) need to be dealt with. In order to achieve accuracy requirements, a great challenge awaits the signal processing practitioner.

References

1. Bentley, J. P., *Principles of Measurement Systems,* 2nd ed., Burnt Mill, UK: Longman Scientific and Technical, 1988.
2. Doebelin, E. O., *Measurement Systems: Application and Design,* 4th ed., Singapore: McGraw-Hill, 1990.
3. Frank, R., *Understanding Smart Sensors,* Boston: Artech House, 1996.
4. Harris, C., *Shock and Vibration Handbook,* 4th ed., McGraw-Hill, 1995.
5. Holman, J. P., *Experimental Methods for Engineers,* 5th ed., Singapore: McGraw-Hill, 1989.

6. Lawrance, A., *Modern Inertial Technology-Navigation, Guidance, and Control,* Springer-Verlag, New York, 1993.
7. McConnell, K. G., *Vibration Testing: Theory and Practice,* New York: Wiley, 1995.
8. *Machine Vibration: Dynamics and Control,* London: Springler, 1992–1996.
9. *Measuring Vibration,* Bruel & Kjaer, 1982.
10. Sydenham, P. H., Hancock, N. H., and Thorn, R., *Introduction to Measurement Science and Engineering,* New York: Wiley, 1989.
11. Tompkins, W. J. and Webster, J. G., *Interfacing Sensors to the IBM PC,* Englewood Cliffs, NJ: Prentice-Hall, 1988.

19.3 Force Measurement

M. A. Elbestawi

Force, which is a vector quantity, can be defined as an action that will cause an acceleration or a certain reaction of a body. This chapter will outline the methods that can be employed to determine the magnitude of these forces.

General Considerations

The determination or measurement of forces must yield to the following considerations: if the forces acting on a body do not produce any acceleration, they must form a *system of forces in equilibrium.* The system is then considered to be in static equilibrium. The forces experienced by a body can be classified into two categories: internal, where the individual particles of a body act on each other, and external otherwise. If a body is supported by other bodies while subject to the action of forces, deformations and/or displacements will be produced at the points of support or contact. The internal forces will be distributed throughout the body until equilibrium is established, and then the body is said to be in a state of tension, compression, or shear. In considering a body at a definite section, it is evident that all the internal forces act in pairs, the two forces being equal and opposite, whereas the external forces act singly.

Hooke's Law

The basis for force measurement results from the physical behavior of a body under external forces. Therefore, it is useful to review briefly the mechanical behavior of materials. When a metal is loaded in uniaxial tension, uniaxial compression, or simple shear (Fig. 19.29), it will behave elastically until a critical value of normal stress (S) or shear stress (τ) is reached, and then it will deform plastically [1]. In the

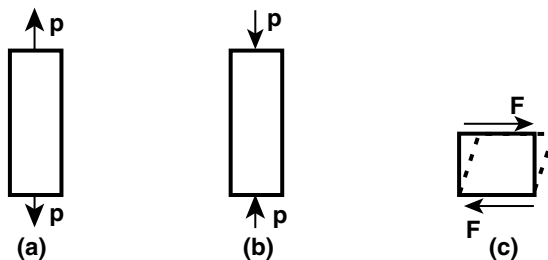



FIGURE 19.29 When a metal is loaded in uniaxial tension (a), uniaxial compression (b), or simple shear(c), it will behave elastically until a critical value of normal stress or shear stress is reached.

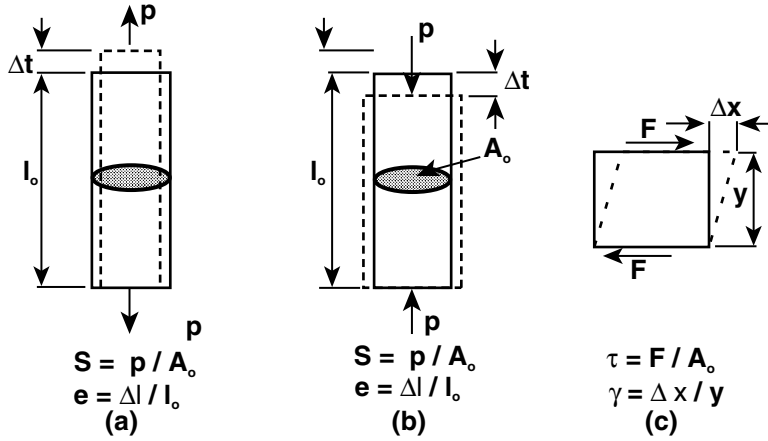



FIGURE 19.30 Elastic stress and strain for: (a) uniaxial tension; (b) uniaxial compression; (c) simple shear [1].

elastic region, the atoms are temporarily displaced but return to their equilibrium positions when the load is removed. Stress (S or τ) and strain (e or γ) in the elastic region are defined as indicated in Fig. 19.30.

$$\nu = -\frac{e_2}{e_1} \tag{19.41}$$

Poisson's ratio (ν) is the ratio of transverse (e_2) to direct (e_1) strain in tension or compression. In the elastic region, ν is between 1/4 and 1/3 for metals. The relation between stress and strain in the elastic region is given by Hooke's law:

$$S = Ee \quad \text{(tension or compression)} \tag{19.42}$$

$$\tau = G\gamma \quad \text{(simple shear)} \tag{19.43}$$

where E and G are the Young's and shear modulus of elasticity, respectively. A small change in specific volume ($\Delta Vol/Vol$) can be related to the elastic deformation, which is shown to be as follows for an isotropic material (same properties in all directions):

$$\frac{\Delta Vol}{Vol} = e_1(1-2\nu) \tag{19.44}$$

The bulk modulus (K = reciprocal of compressibility) is defined as follows:

$$K = \Delta p \Big/ \left(\frac{\Delta Vol}{Vol}\right) \tag{19.45}$$

where Δp is the pressure acting at a particular point. For an elastic solid loaded in uniaxial compression (S):

$$K = S \Big/ \left(\frac{\Delta Vol}{Vol}\right) = \frac{S}{e_1(1-2\nu)} = \frac{E}{1-2\nu} \tag{19.46}$$

Thus, an elastic solid is compressible as long as ν is less than 1/2, which is normally the case for metals. Hooke's law, Eq. (19.42), for uniaxial tension can be generalized for a three-dimensional elastic condition.

The theory of elasticity is well established and is used as a basis for force measuring techniques. Note that the measurement of forces in separate engineering applications is very application specific, and care must be taken in the selection of the measuring techniques outlined below.

Basic Methods of Force Measurement

An unknown force may be measured by the following means:

1. balancing the unknown force against a standard mass through a system of levers,
2. measuring the acceleration of a known mass,
3. equalizing it to a magnetic force generated by the interaction of a current-carrying coil and a magnet,
4. distributing the force on a specific area to generate pressure and then measuring the pressure,
5. converting the applied force into the deformation of an elastic element.

The aforementioned methods used for measuring forces yield a variety of designs of measuring equipment. The challenge involved with the task of measuring force resides primarily in sensor design. The basics of sensor design can be resolved into two problems:

1. primary geometric, or physical constraints, governed by the application of the force sensor device;
2. the means by which the force can be converted into a workable signal form (such as electronic signals or graduated displacements).

The remaining sections will discuss the types of devices used for force-to-signal conversion and finally illustrate some examples of applications of these devices for measuring forces.

Force Sensors

Force sensors are required for a basic understanding of the response of a system. For example, cutting forces generated by a machining process can be monitored to detect a tool failure or to diagnose the causes of this failure in controlling the process parameters, and in evaluating the quality of the surface produced. Force sensors are used to monitor impact forces in the automotive industry. Robotic handling and assembly tasks are controlled by detecting the forces generated at the end effector. Direct measurement of forces is useful in controlling many mechanical systems.

Some types of force sensors are based on measuring a deflection caused by the force. Relatively high deflections (typically, several micrometers) would be necessary for this technique to be feasible. The excellent elastic properties of helical springs make it possible to apply them successfully as force sensors that transform the load to be measured into a deflection. The relation between force and deflection in the elastic region is demonstrated by Hooke's law. Force sensors that employ strain gage elements or piezoelectric (quartz) crystals with built-in microelectronics are common. Both impulsive forces and slowly varying forces can be monitored using these sensors.

Of the available force measuring techniques, a general subgroup can be defined as that of load cells. Load cells are comprised generally of a rigid outer structure, some medium that is used for measuring the applied force, and the measuring gage. Load cells are used for sensing large, static, or slowly varying forces with little deflection and are a relatively accurate means of sensing forces. Typical accuracies are of the order of 0.1% of the full-scale readings. Various strategies can be employed for measuring forces that are strongly dependent on the design of the load cell. For example, Fig. 19.31 illustrates different types of load cells that can be employed in sensing large forces for relatively little cost. The hydraulic load cell employs a very stiff outer structure with an internal cavity filled with a fluid. Application of a load increases the oil pressure, which can be read off an accurate gage.

Other sensing techniques can be utilized to monitor forces, such as piezoelectric transducers for quicker response of varying loads, pneumatic methods, strain gages, etc. The proper sensing technique needs special consideration based on the conditions required for monitoring.

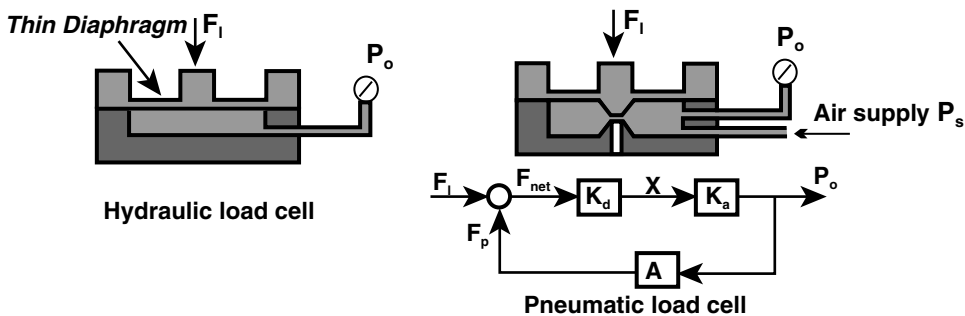



FIGURE 19.31 Different types of load cells [2].

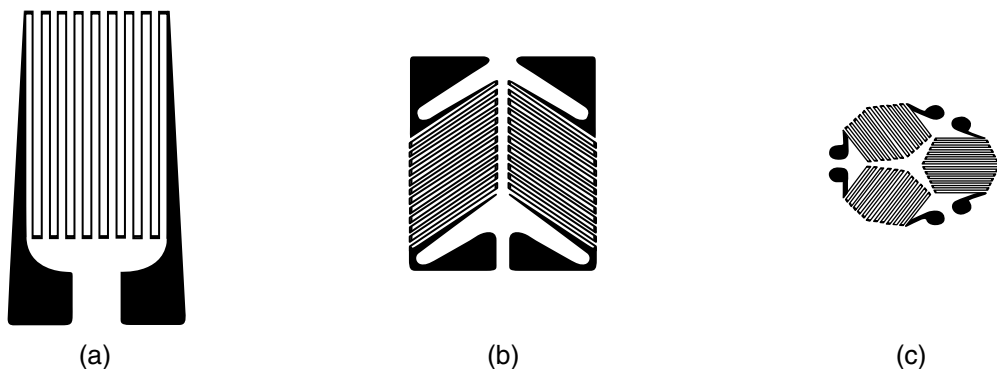



FIGURE 19.32 Configuration of metal-foil resistance strain gages: (a) single element, (b) two element, and (c) three element.

Strain Gage Load Cell

The strain gage load cell consists of a structure that elastically deforms when subjected to a force and a strain gage network that produces an electrical signal proportional to this deformation. Examples of this are beam and ring types of load cells.

Strain Gages

Strain gages use a length of gage wire to produce the desired resistance (which is usually about 120 Ω) in the form of a flat coil. This coil is then cemented (bonded) between two thin insulating sheets of paper or plastic. Such a gage cannot be used directly to measure deflection. It has to be first fixed properly to a member to be strained. After bonding the gage to the member, they are baked at about 195°F (90°C) to remove moisture. Coating the unit with wax or resin will provide some mechanical protection. The resistance between the member under test and the gage itself must be at least 50 MΩ. The total area of all conductors must remain small so that the cement can easily transmit the force necessary to deform the wire. As the member is stressed, the resulting strain deforms the strain gage and the cross-sectional area diminishes. This causes an increase in resistivity of the gage that is easily determined. In order to measure very small strains, it is necessary to measure small changes of the resistance per unit resistance ($\Delta R/R$). The change in the resistance of a bonded strain gage is usually less than 0.5%. A wide variety of gage sizes and grid shapes are available, and typical examples are shown in Fig. 19.32.

The use of strain gages to measure force requires careful consideration with respect to rigidity and environment. By virtue of their design, strain gages of shorter length generally possess higher response frequencies (examples: 660 kHz for a gage of 0.2 mm and 20 kHz for a gage of 60 mm in length). The environmental considerations focus mainly on the temperature of the gage. It is well known that resistance is a function of temperature and, thus, strain gages are susceptible to variations in temperature.

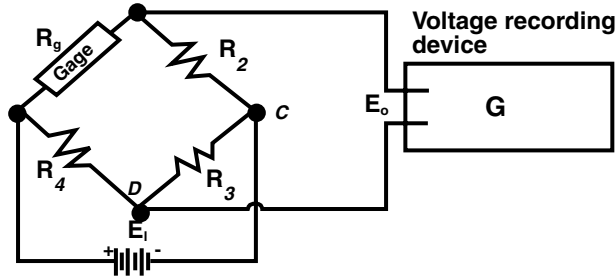



FIGURE 19.33 The Wheatstone bridge.

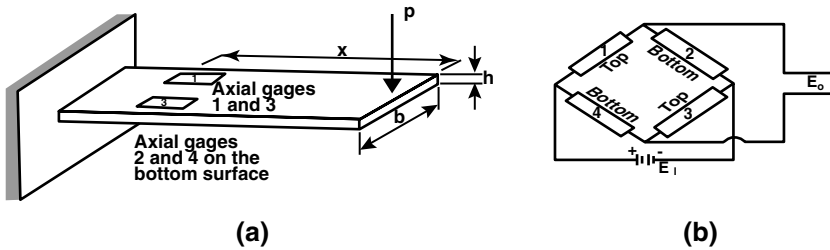



FIGURE 19.34 Beam-type load cells: (a) a selection of beam-type load cells (elastic element with strain gages), and (b) gage positions in the Wheatstone bridge [3].

Thus, if it is known that the temperature of the gage will vary due to any influence, temperature compensation is required in order to ensure that the force measurement is accurate.

A Wheatstone bridge (Fig. 19.33) is usually used to measure this small order of magnitude. In Fig. 19.33, no current will flow through the galvanometer (G) if the four resistances satisfy a certain condition. In order to demonstrate how a Wheatstone bridge operates [3], a voltage scale has been drawn at points C and D of Fig. 19.33. Assume that R_1 is a bonded gage and that initially Eq. (19.47) is satisfied. If R_1 is now stretched so that its resistance increases by one unit ($+\Delta R$), the voltage at point D will be increased from zero to plus one unit of voltage ($+\Delta V$), and there will be a voltage difference of one unit between C and D that will give rise to a current through C. If R_4 is also a bonded gage, and at the same time that R_1 changes by $+\Delta R$, R_4 changes by $-\Delta R$, the voltage at D will move to $+2\Delta V$. Also, if at the same time, R_2 changes by $-\Delta R$, and R_3 changes by $+\Delta R$, then the voltage of point C will move to $-2\Delta V$, and the voltage difference between C and D will now be $4\Delta V$. It is then apparent that although a single gage can be used, the sensitivity can be increased fourfold if two gages are used in tension while two others are used in compression.

$$\frac{R_1}{R_4} = \frac{R_2}{R_3} \tag{19.47}$$

The grid configuration of the metal-foil resistance strain gages is formed by a photo-etching process. The shortest gage available is 0.20 mm; the longest is 102 mm. Standard gage resistances are 120 Ω and 350 Ω. A strain gage exhibits a resistance change $\Delta R/R$ that is related to the strain in the direction of the grid lines by the expression in Eq. (19.48) (where S_g is the gage factor or calibration constant for the gage).

$$\frac{\Delta R}{R} = S_g \varepsilon \tag{19.48}$$

Beam-Type Load Cell

Beam-type load cells are commonly employed for measuring low-level loads [3]. A simple cantilever beam (see Fig. 19.34(a)) with four strain gages, two on the top surface and two on the bottom surface (all oriented along the axis of the beam) is used as the elastic member (sensor) for the load cell. The gages

are wired into a Wheatstone bridge as shown in Fig. 19.34(b). The load P produces a moment $M = Px$ at the gage location (x) that results in the following strains:

$$\varepsilon_1 = -\varepsilon_2 = \varepsilon_3 = -\varepsilon_4 = \frac{6M}{Ebh^2} = \frac{6Px}{Ebh^2} \tag{19.49}$$

where b is the width of the cross-section of the beam and h is the height of the cross-section of the beam. Thus, the response of the strain gages is obtained from Eq. (19.50).

$$\frac{\Delta R_1}{R_1} = -\frac{\Delta R_2}{R_2} = \frac{\Delta R_3}{R_3} = -\frac{\Delta R_4}{R_4} = \frac{6S_gPx}{Ebh^2} \tag{19.50}$$

The output voltage E_o from the Wheatstone bridge, resulting from application of the load P, is obtained from Eq. (19.51). If the four strain gages on the beam are assumed to be identical, then Eq. (19.51) holds.

$$E_o = \frac{6S_gPxE_1}{Ebh^2} \tag{19.51}$$

The range and sensitivity of a beam-type load cell depends on the shape of the cross-section of the beam, the location of the point of application of the load, and the fatigue strength of the material from which the beam is fabricated.

Ring-Type Load Cell

Ring-type load cells incorporate a proving ring (see Fig. 19.35) as the elastic element. The ring element can be designed to cover a very wide range of loads by varying the diameter D, the thickness t, or the depth w of the ring. Either strain gages or a linear variable-differential transformer (LVDT) can be used as the sensor.

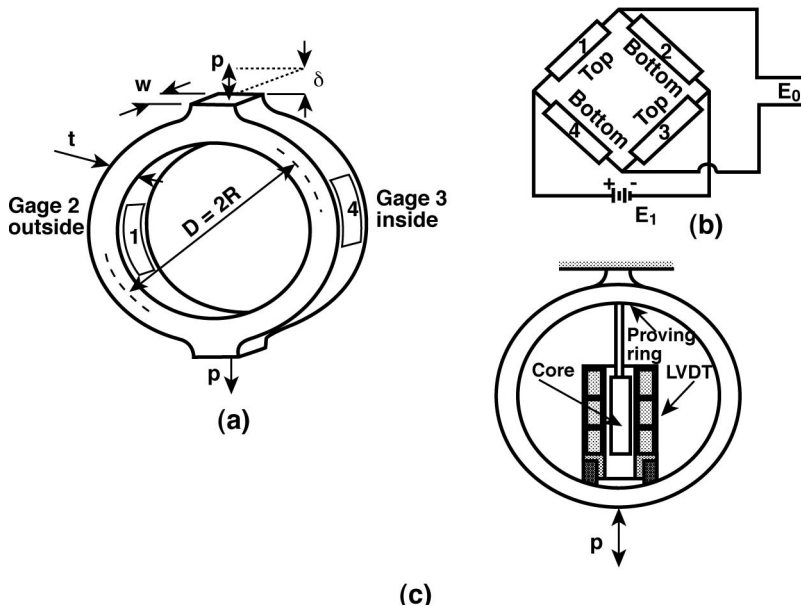



FIGURE 19.35 Ring-type load cells: (*a*) elastic element with strain-gage sensors; (*b*) gage positions in the Wheatstone bridge; and (*c*) elastic element with an LVDT sensor [3].

The load P is linearly proportional to the output voltage E_o. The sensitivity of the ring-type load cell with an LVDT sensor depends on the geometry of the ring (R, t, and w), the material from which the ring is fabricated (E), and the characteristics of the LVDT (S and E_i). The range of a ring-type load cell is controlled by the strength of the material used in fabricating the ring.

Piezoelectric Methods

A piezoelectric material exhibits a phenomenon known as the *piezoelectric effect*. This effect states that when asymmetrical, elastic crystals are deformed by a force, an electrical potential will be developed within the distorted crystal lattice. This effect is reversible. That is, if a potential is applied between the surfaces of the crystal, it will change its physical dimensions [4]. Elements exhibiting piezoelectric qualities are sometimes known as electrorestrictive elements.

The magnitude and polarity of the induced surface charges are proportional to the magnitude and direction of the applied force [4]:

$$Q = dF \tag{19.52}$$

where d is the charge sensitivity (a constant for a given crystal) of the crystal in C/N. The force F causes a thickness variation Δt meters of the crystal:

$$F = \frac{aY}{t}\Delta t \tag{19.53}$$

where a is area of crystal, t is thickness of crystal, and Y is Young's modulus.

$$Y = \frac{\text{stress}}{\text{strain}} = \frac{Ft}{a\Delta t} \tag{19.54}$$

The charge at the electrodes gives rise to a voltage $E_o = Q/C$, where C is capacitance in farads between the electrodes and $C = \varepsilon a/t$ where ε is the absolute permittivity.

$$E_o = \frac{dF}{C} = \frac{dtF}{\varepsilon a} \tag{19.55}$$

The voltage sensitivity $= g = d/\varepsilon$ in volt meter per newton can be obtained as:

$$E_o = g\frac{t}{a}F = gtP \tag{19.56}$$

The piezoelectric materials used are quartz, tourmaline, Rochelle salt, ammonium dihydrogen phosphate (ADP), lithium sulfate, barium titanate, and lead zirconate titanate (PZT) [4]. Quartz and other earthly piezoelectric crystals are naturally polarized. However, synthetic piezoelectric material, such as barium titanate ceramic, are made by baking small crystallites under pressure and then placing the resultant material in a strong dc electric field [4]. After that, the crystal is polarized, along the axis on which the force will be applied, to exhibit piezoelectric properties. Artificial piezoelectric elements are free from the limitations imposed by the crystal structure and can be molded into any size and shape. The direction of polarization is designated during their production process.

The different modes of operation of a piezoelectric device for a simple plate are shown in Fig. 19.36 [4]. By adhering two crystals together so that their electrical axes are perpendicular, bending moments or torque can be applied to the piezoelectric transducer and a voltage output can be produced (Fig. 19.37) [4]. The range of forces that can be measured using piezoelectric transducers are from 1 to 200 kN and at a ratio of 2×10^5.

Piezoelectric crystals can also be used in measuring an instantaneous change in the force (dynamic forces). A thin plate of quartz can be used as an electronic oscillator. The frequency of these oscillations will be dominated by the natural frequency of the thin plate. Any distortion in the shape of the plate

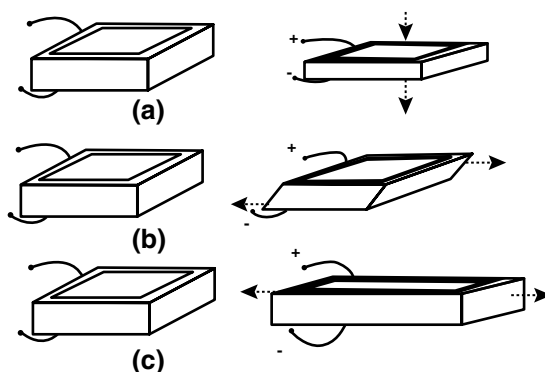



FIGURE 19.36 Modes of operation for a simple plate as a piezoelectric device [4].

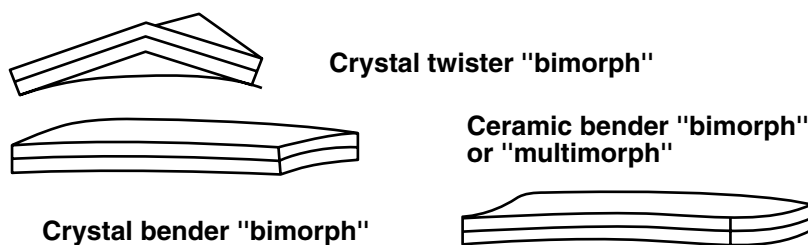



FIGURE 19.37 Curvature of "twister" and "bender" piezoelectric transducers when voltage applied [4].

caused by an external force, alters the oscillation frequency. Hence, a dynamic force can be measured by the change in frequency of the oscillator.

Resistive Method

The resistive method employs the fact that when the multiple contact area between semiconducting particles (usually carbon) and the distance between the particles are changed, the total resistance is altered. The design of such transducers yields a very small displacement when a force is applied. A transducer might consist of 2–60 thin carbon disks mounted between a fixed and a movable electrode. When a force is applied to the movable electrode and the carbon disks move together by 5–250 μm per interface, the transfer function of their resistance against the applied force is approximately hyperbolic, that is, highly nonlinear. The device is also subject to large hysteresis and drift together with a high transverse sensitivity.

In order to reduce hysteresis and drift, rings are used instead of disks. The rings are mounted on an insulated rigid core and prestressed. This almost completely eliminates any transverse sensitivity error. The core's resonant frequency is high and can occur at a frequency as high as 10 kHz. The possible measuring range of such a transducer is from 0.1 to 10 kg. The accuracy and linear sensitivity of this transducer is very poor.

Inductive Method

The inductive method utilizes the fact that a change in mechanical stress of a ferromagnetic material causes its permeability to alter. The changes in magnetic flux are converted into induced voltages in the pickup coils as the movement takes place. This phenomenon is known as the *Villari effect* or *magnetostriction*. It is known to be particularly strong in nickel–iron alloys.

Transducers utilizing the Villari effect consist of a coil wound on a core of magnetostrictive material. The force to be measured is applied on this core, stressing it and causing a change in its permeability and inductance. This change can be monitored and used for determining the force.

The applicable range for this type of transducer is a function of the cross-sectional area of the core. The accuracy of the device is determined by a calibration process. This transducer has poor linearity and is subject to hysteresis. The permeability of a magnetostrictive material increases when it is subject to pure torsion, regardless of direction. A flat frequency response is obtained over a wide range from 150 to 15,000 Hz.

Piezotransistor Method

Devices that utilize *anisotropic stress effects* are described as piezotransistors. In this effect, if the upper surface of a *p-n* diode is subjected to a localized stress, a significant reversible change occurs in the current across the junction. These transistors are usually silicon nonplanar type, with an emitter base junction. This junction is mechanically connected to a diaphragm positioned on the upper surface of a typical TO-type can [4]. When a pressure or a force is applied to the diaphragm, an electronic charge is produced. It is advisable to use these force-measuring devices at a constant temperature by virtue of the fact that semiconducting materials also change their electric properties with temperature variations. The attractive characteristic of piezotransistors is that they can withstand a 500% overload.

Multicomponent Dynamometers Using Quartz Crystals as Sensing Elements

The Piezoelectric Effects in Quartz

For force measurements, the *direct piezoelectric effect* is utilized. The direct longitudinal effect measures compressive force; the direct shear effect measures shear force in one direction. For example, if a disk of crystalline quartz (SiO_2) cut normally to the crystallographic x-axis is loaded by a compression force, it will yield an electric charge, nominally 2.26 pC/N. If a disk of crystalline quartz is cut normally to the crystallographic y-axis, it will yield an electric charge (4.52 pC/N) if loaded by a shear force in one specific direction. Forces applied in the other directions will not generate any output [5].

A charge amplifier is used to convert the charge yielded by a quartz crystal element into a proportional voltage. The range of a charge amplifier with respect to its conversion factor is determined by a feedback capacitor. Adjustment to mechanical units is obtained by additional operational amplifiers with variable gain.

The Design of Quartz Multicomponent Dynamometers

The main element for designing multicomponent dynamometers is the three-component force transducer (Fig. 19.38). It contains a pair of X-cut quartz disks for the normal force component and a pair of Y-cut quartz disks (shear-sensitive) for each shear force component.

Three-component dynamometers can be used for measuring cutting forces during machining. Four three-component force transducers sandwiched between a base plate and a top plate are shown in Fig. 19.38. The force transducer is subjected to a preload as shear forces are transmitted by friction. The four force transducers experience a drastic change in their load, depending on the type and position of force application. An overhanging introduction of the force develops a tensile force for some transducers, thus reducing the preload. Bending of the dynamometer top plate causes bending and shearing stresses. The measuring ranges of a dynamometer depend not only on the individual forces, but also on the individul bending stresses.

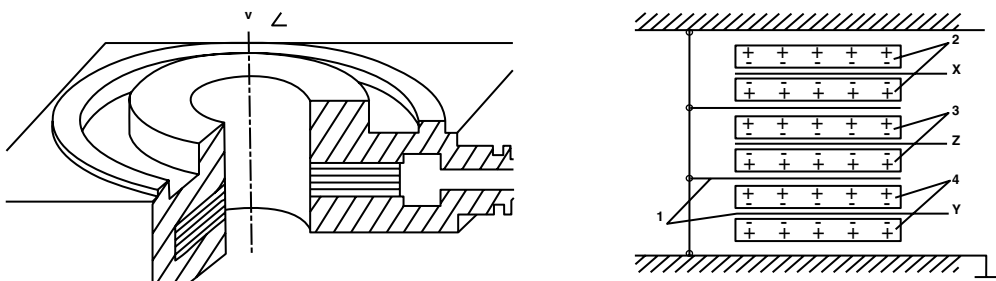



FIGURE 19.38 Three-component force transducer.

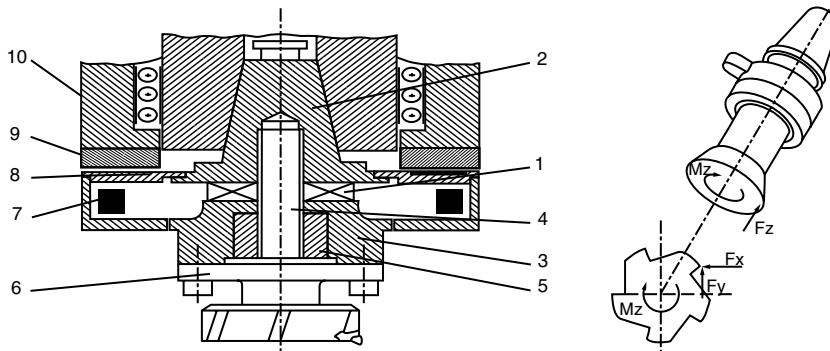



FIGURE 19.39 Force measuring system to determine the tool-related cutting forces in five-axis milling [6].

Measuring Signals Transmitted by Telemetry

Figure 19.39 shows the newly designed force measuring system RCD (rotating cutting force dynamometer). A ring-shaped sensor (1) is fitted in a steep angle taper socket (2) and a base ring (3) allowing sensing of the three force components F_x, F_y, and F_z at the cutting edge as well as the moment M_z. The physical operating principle of this measuring cell is based on the piezoelectric effect in quartz plates. The quartz plates incorporated in the sensor are aligned so that the maximum cross-sensitivity between the force components is 1%. As a result of the rigid design of the sensor, the resonant frequencies of the force measuring system range from 1200 to 3000 Hz and the measuring ranges cover a maximum of 10 kN [6].

Force-proportional charges produced at the surfaces of the quartz plates are converted into voltages by four miniature charge amplifiers (7) in hybrid construction. These signals are then filtered by specific electrical circuitry to prevent aliasing effects, and digitized with 8-bit resolution using a high sampling rate (pulse-code modulation). The digitized signals are transmitted by a telemetric unit consisting of a receiver and transmitter module, an antenna at the top of the rotating force measuring system (8), as well as a fixed antenna (9) on the splash cover of the two-axis milling head (10). The electrical components, charge amplifier, and transmitter module are mounted on the circumference of the force measuring system [6].

The cutting forces and the moment measured are digitized with the force measuring system described above. They are modulated on an FM carrier and transmitted by the rotating transmitter to the stationary receiver. The signals transmitted are fed to an external measured-variable conditioning unit.

Measuring Dynamic Forces

Any mechanical system can be considered in the first approximation as a weakly damped oscillator consisting of a spring and a mass. If a mechanical system has more than one resonant frequency, the lowest one must be taken into consideration. As long as the test frequency remains below 10% of the resonant frequency of the reference transducer (used for calibration), the difference between the dynamic sensitivity obtained from static calibration will be less than 1%. The above considerations assume a sinusoidal force signal. The static calibration of a reference transducer is also valid for dynamic calibration purposes if the test frequency is much lower (at least 10 times lower) than the resonant frequency of the system.

Capacitive Force Transducer

A transducer that uses capacitance variation can be used to measure force. The force is directed onto a membrane whose elastic deflection is detected by a capacitance variation. A highly sensitive force transducer can be constructed because the capacitive transducer senses very small deflections accurately. An electronic circuit converts the capacitance variations into DC-voltage variations [7].

A capacitance sensor consists of two metal plates separated by an air gap. The capacitance C between terminals is given by the expression:

$$C = \varepsilon_o \varepsilon_r \frac{A}{h} \tag{19.57}$$

where

C = capacitance in farads (F),
ε_o = dielectric constant of free space,
ε_r = relative dielectric constant of the insulator,
A = overlapping area for the two plates,
h = thickness of the gap between the two plates.

The sensitivity of capacitance-type sensors is inherently low. Theoretically, decreasing the gap h should increase the sensitivity; however, there are practical electrical and mechanical conditions that preclude high sensitivities. One of the main advantages of the capacitive transducer is that moving of one of its plates relative to the other requires an extremely small force to be applied. A second advantage is stability and the sensitivity of the sensor is not influenced by pressure or temperature of the environment.

Force Sensing Resistors (Conductive Polymers)

Force sensing resistors (FSRs) utilize the fact that certain polymer thick-film devices exhibit decreasing resistance with the increase of an applied force. A force sensing resistor is made up of two parts. The first is a resistive material applied to a film. The second is a set of digitating contacts applied to another film. Figure 19.40 shows this configuration. The resistive material completes the electrical circuit between the two sets of conductors on the other film. When a force is applied to this sensor, a better connection is made between the contacts; hence, the conductivity is increased. Over a wide range of forces, it turns out that the conductivity is approximately a linear function of force. Figure 19.41 shows the resistance of the sensor as a function of force. It is important to note that there are three possible regions for the sensor to operate. The first abrupt transition occurs somewhere in the vicinity of 10 g of force. In this region, the resistance changes very rapidly. This behavior is useful when one is designing switches using force sensing resistors.

FSRs should not be used for accurate measurements of force because sensor parts may exhibit 15–25% variation in resistance between each other. However, FSRs exhibit little hysteresis and are considered far less costly than other sensing devices. Compared to piezofilm, the FSR is far less sensitive to vibration and heat.

Magnetoresistive Force Sensors

The principle of *magnetoresistive force sensors* is based on the fact that metals, when cooled to low temperatures, show a change of resistivity when subjected to an applied magnetic field. Bismuth, in particular, is quite sensitive in this respect. In practice, these devices are severely limited because of their high sensitivity to ambient temperature changes.

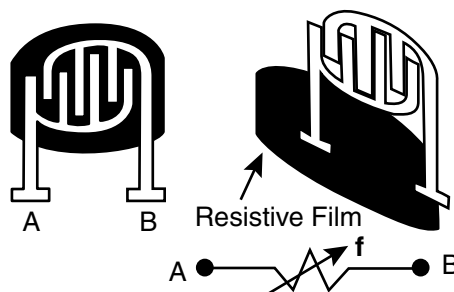



FIGURE 19.40 Diagram of a typical force sensing resistor (FSR).

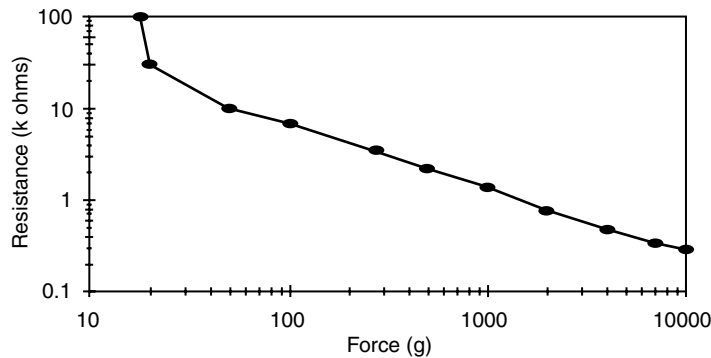



FIGURE 19.41 Resistance as a function of force for a typical force sensing resistor.

Magnetoelastic Force Sensors

Magnetoelastic transducer devices operate based on the Joule effect, that is, a ferromagnetic material is dimensionally altered when subjected to a magnetic field. The principle of operation is as follows: Initially, a current pulse is applied to the conductor within the waveguide. This sets up a magnetic field circumference-wise around the waveguide over its entire length. There is another magnetic field generated by the permanent magnet that exists only where the magnet is located. This field has a longitudinal component. These two fields join vectorally to form a helical field near the magnet which, in turn, causes the waveguide to experience a minute torsional strain or twist only at the location of the magnet. This twist effect is known as the *Wiedemann effect* [8].

Magnetoelastic force transducers have a high frequency response (on the order of 20 kHz). Some of the materials that exhibit magnetoelastic include Monel metal, Permalloy, Cekas, Alfer, and a number of nickel–iron alloys. Disadvantages of these transducers include: (1) the fact that excessive stress and aging may cause permanent changes, (2) zero drift and sensitivity changes due to temperature sensitivity, and (3) hysteresis errors.

Torsional Balances

Balancing devices that utilize the deflection of a spring may also be used to determine forces. *Torsional balances* are equal arm-scale-force measuring devices. They are comprised of horizontal steel bands instead of pivots and bearings. The principle of operation is based on force application on one of the arms that will deflect the torsional spring (within its design limits) in proportion to the applied force. This type of instrument is susceptible to hysteresis and temperature errors and, therefore, is not used for precise measurements.

Tactile Sensors

Tactile sensors are usually interpreted as a touch sensing technique. Tactile sensors cannot be considered as simple touch sensors, where very few discrete force measurements are made. In tactile sensing, a force "distribution" is measured using a closely spaced array of force sensors.

Tactile sensing is important in both grasping and object identification operations. Grasping an object must be done in a stable manner so that the object is not allowed to slip or get damaged. Object identification includes recognizing the shape, location, and orientation of a product, as well as identifying surface properties and defects. Ideally, these tasks would require two types of sensing [9]:

1. continuous sensing of contact forces,
2. sensing of the surface deformation profile.

These two types of data are generally related through stress–strain relations of the tactile sensor. As a result, almost continuous variable sensing of tactile forces (the sensing of the tactile deflection profile) is achieved.

Tactile Sensor Requirements

Significant advances in tactile sensing are taking place in the robotics area. Applications include automated inspection of surface profiles, material handling or parts transfer, parts assembly, and parts identification and gaging in manufacturing applications and fine-manipulation tasks. Some of these applications may need only simple touch (force–torque) sensing if the parts being grasped are properly oriented and if adequate information about the process is already available.

Naturally, the main design objective for tactile sensing devices has been to mimic the capabilities of human fingers [9]. Typical specifications for an industrial tactile sensor include:

1. Spatial resolution of about 2 mm
2. Force resolution (sensitivity) of about 2 g
3. Maximum touch force of about 1 kg
4. Low response time of 5 ms
5. Low hysteresis
6. Durability under extremely difficult working conditions
7. Insensitivity to change in environmental conditions (temperature, dust, humidity, vibration, etc.)
8. Ability to monitor slip

Tactile Array Sensor

Tactile array sensors (Fig. 19.42) consist of a regular pattern of sensing elements to measure the distribution of pressure across the fingertip of a robot. The 8 × 8 array of elements at 2 mm spacing in each direction provides 64 force sensitive elements. Table 19.2 outlines some of the characteristics of early tactile array sensors. The sensor is composed of two crossed layers of copper strips separated by strips of thin silicone rubber. The sensor forms a thin, compliant layer that can be easily attached to a variety of fingertip shapes and sizes. The entire array is sampled by computer.

TABLE 19.2 Summary of Some of the Characteristics of Early Tactile Array Sensors

Device Parameter	Size of Array (4 × 4)	(8 × 8)	(16 × 16)
Cell spacing (mm)	4.00	2.00	1.00
Zero-pressure capacitance (fF)	6.48	1.62	0.40
Rupture force (N)	18.90	1.88	0.19
Max. linear capacitance (fF)	4.80	1.20	0.30
Max. output voltage (V)	1.20	0.60	0.30
Max. resolution (bit)	9.00	8.00	8.00
Readout (access) time (μs)	—	<20	—

©IEEE 1985.

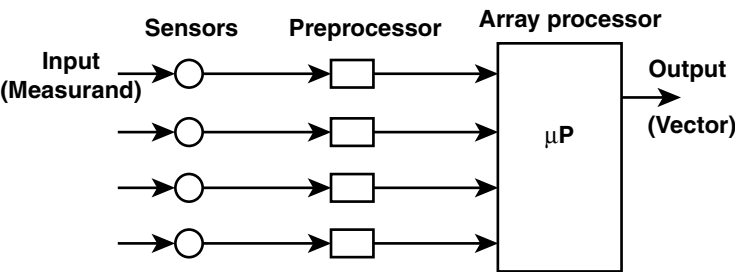



FIGURE 19.42 Tactile array sensor.

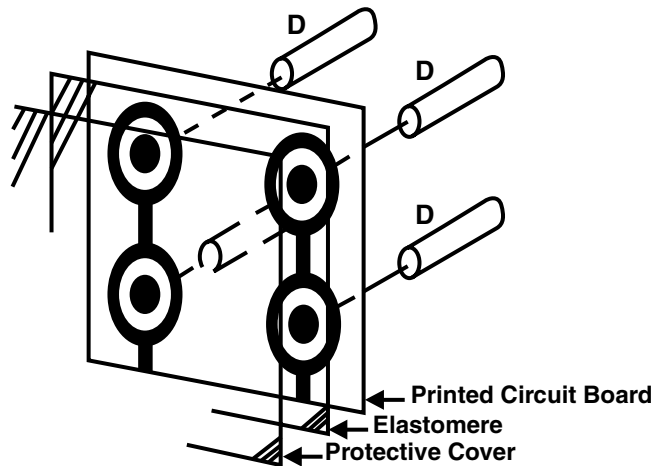



FIGURE 19.43 Typical taxel sensor array.

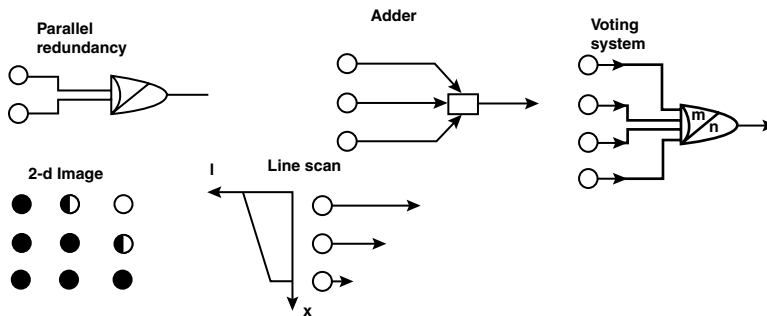



FIGURE 19.44 General arrangement of an intelligent sensor array system [9].

A typical tactile sensor array can consist of several sensing elements. Each element or taxel (Fig. 19.43) is used to sense the forces present. Since tactile sensors are implemented in applications where sensitivity providing semblance to human touch is desired, an elastomer is utilized to mimic the human skin. The elastomer is generally a conductive material whose electrical conductivity changes locally when pressure is applied. The sensor itself consists of three layers: a protective covering, a sheet of conductive elastomer, and a printed circuit board. The printed circuit board consists of two rows of two "bullseyes," each with conductive inner and outer rings that compromise the taxels of the sensor. The outer rings are connected together and to a column-select transistor. The inner rings are connected to diodes (D) in Fig. 19.43. Once the column in the array is selected, the current flows through the diodes, through the elastomer, and thence through a transistor to ground. As such, it is generally not possible to excite just one taxel because the pressure applied causes a local deformation in neighboring taxels. This situation is called *crosstalk* and is eliminated by the diodes [10].

Tactile array sensor signals are used to provide information about the contact kinematics. Several feature parameters, such as contact location, object shape, and the pressure distribution, can be obtained. The general layout of a sensor array system can be seen in Fig. 19.44. An example of this is a contact and force sensing finger. This tactile finger has four contact sensors made of piezoelectric polymer strips on the surface of the fingertip that provide dynamic contact information. A strain gage force sensor provides static grasp force information.

References

1. Shaw, M. C., *Metal Cutting Principles,* Oxford: Oxford Science Publications, Clarendon Press, 1989.
2. Doebelin, E. O., *Measurement Systems, Application and Design,* 4th ed., New York: McGraw-Hill, 1990.
3. Dally, J. W., Riley, W. F., and McConnel, K. G., *Instrumentation for Engineering Measurements,* New York: John Wiley & Sons, 1984.
4. Mansfield, P. H., *Electrical Transducers for Industrial Measurement,* London: The Butterworth Group, 1973.
5. Martini, K. H., Multicomponent dynamometers using quartz crystals as sensing elements, *ISA Trans.,* 22(1), 1983.
6. Spur, G., Al-Badrawy, S. J., and Stirnimann, J., Measuring the Cutting Force in Five-Axis Milling, Translated paper "Zerpankraftmessung bei der funfachsigen Frasbearbeitung," Zeitschrift fur wirtschaftliche Fertigung und Automatisierung 9/93 Carl Hanser, Munchen, Kistler Piezo-Instrumentation, 20.162e 9.94.
7. Nachtigal, C. L., *Instrumentation and Control, Fundamentals and Applications,* Wiley Series in Mechanical Engineering Practice, New York: Wiley Interscience, John Wiley & Sons, 1990.
8. DeSilva, C. W., *Control Sensors and Actuators,* Englewood Cliffs, NJ: Prentice-Hall, 1989.
9. Gardner, J. W., *Microsensors Principles and Applications,* New York: John Wiley & Sons, 1995.
10. Stadler, W., *Analytical Robotics and Mechatronics*, New York: McGraw-Hill, 1995.

Further Information

Wright, C. P., *Applied Measurement Engineering, How to Design Effective Mechanical Measurement Systems,* Englewood Cliffs, NJ: Prentice Hall, 1995.

Herceg, E. E., *Handbook of Measurement and Control,* Pennsauken, NJ: Schavitz Engineering, 1972.

Considine, D. M., *Encyclopedia of Instrumentation and Control,* New York: McGraw-Hill, 1971.

Norton, H. N., *Sensor and Analyzer Handbook,* Englewood Cliffs, NJ: Prentice-Hall, 1982.

Sze, S. M., *Semiconductor Sensors,* New York: John Wiley & Sons, 1994

Lindberg, B., and Lindstrom, B., Measurements of the segmentation frequency in the chip formation process, *Ann. CIRP,* 32(1), 1983.

Tlusty, J., and Andrews, G. C., A critical review of sensors for unmanned machinning, *Ann. CIRP,* 32(2), 1983.

19.4 Torque and Power Measurement

Ivan J. Garshelis

Torque, speed, and power are the defining mechanical variables associated with the functional performance of rotating machinery. The ability to accurately measure these quantities is essential for determining a machine's efficiency and for establishing operating regimes that are both safe and conducive to long and reliable services. Online measurements of these quantities enable real-time control, help ensure consistency in product quality, and can provide early indications of impending problems. Torque and power measurements are used in testing advanced designs of new machines and in the development of new machine components. Torque measurements also provide a well-established basis for controlling and verifying the tightness of many types of threaded fasteners. This chapter describes the basic concepts as well as the various methods and apparati in current use for the measurement of torque and power; the measurement of speed, or more precisely, angular velocity, is discussed elsewhere [1].

Fundamental Concepts

Angular Displacement, Velocity, and Acceleration

The concept of *rotational* motion is readily formalized: all points within a rotating rigid body move in parallel or coincident planes while remaining at fixed distances from a line called the *axis*. In a perfectly rigid body, all points also remain at fixed distances from each other. Rotation is perceived as a change in the angular position of a reference point on the body, i.e., as its *angular displacement*, $\Delta\theta$, over some time interval, Δt. The motion of that point, and therefore of the whole body, is characterized by its clockwise (CW) or counterclockwise (CCW) *direction* and by its *angular velocity*, $\omega = \Delta\theta/\Delta t$. If during a time interval Δt, the velocity changes by $\Delta\omega$, the body is undergoing an *angular acceleration*, $\alpha = \Delta\omega/\Delta t$. With angles measured in radians, and time in seconds, units of ω become radians per second (rad s^{-1}) and of α, radians per second per second (rad s^{-2}). Angular velocity is often referred to as *rotational speed* and measured in numbers of complete revolutions per minute (rpm) or per second (rps).

Force, Torque, and Equilibrium

Rotational motion, as with motion in general, is controlled by *forces* in accordance with Newton's laws. Because a force directly affects only that component of motion in its line of action, forces or components of forces acting in any plane that includes the axis produce no tendency for rotation about that axis. Rotation can be initiated, altered in velocity, or terminated only by a *tangential force* F_t acting at a finite radial distance l from the axis. The effectiveness of such forces increases with both F_t and l; hence, their product, called a *moment*, is the activating quantity for rotational motion. A moment about the rotational axis constitutes a *torque*. Figure 19.45(a) shows a force F acting at an angle β to the tangent at a point P, distant l (the moment arm) from the axis. The torque T is found from the *tangential component* of F as

$$T = F_t l = (F \cos \beta) l \tag{19.58}$$

The combined effect, known as the *resultant*, of any number of torques acting at different locations along a body is found from their *algebraic sum*, wherein torques tending to cause rotation in CW and CCW directions are assigned opposite signs. Forces, hence torques, arise from physical contact with other solid bodies, motional interaction with fluids, or via gravitational (including inertial), electric, or magnetic force fields. The *source* of each such torque is subjected to an equal, but oppositely directed, *reaction* torque. With force measured in newtons and distance in meters, Eq. (19.58) shows the unit of torque to be a Newton meter (N m).

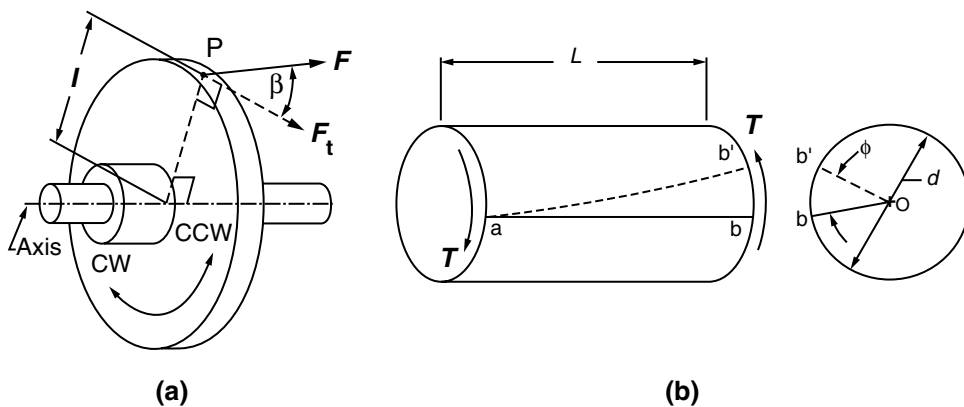



FIGURE 19.45 (a) The off-axis force F at P produces a torque $T = (F \cos \beta)l$ tending to rotate the body in the CW direction. (b) Transmitting torque T over length L twists the shaft through angle ϕ.

A nonzero resultant torque will cause the body to undergo a proportional angular acceleration, found, by application of Newton's second law, from

$$T_r = I\alpha \tag{19.59}$$

where I, having units of kilogram square meter (kg m^2), is the moment of inertia of the body around the axis (i.e., its *polar* moment of inertia). Equation (19.59) is applicable to any body regardless of its state of motion. When $\alpha = 0$, Eq. (19.59) shows that T_r is also zero; the body is said to be in *equilibrium*. For a body to be in equilibrium, there must be either more than one *applied* torque, or none at all.

Stress, Rigidity, and Strain

Any portion of a rigid body in equilibrium is also in equilibrium; hence, as a condition for equilibrium of the portion, any torques applied thereto from *external* sources must be balanced by equal and directionally opposite *internal* torques from adjoining portions of the body. Internal torques are *transmitted* between adjoining portions by the collective action of *stresses* over their common cross-sections. In a solid body having a round cross-section (e.g., a typical shaft), the *shear stress* τ varies linearly from zero at the axis to a maximum value at the surface. The shear stress, τ_m, at the surface of a shaft of diameter, d, transmitting a torque, T, is found from

$$\tau_m = \frac{16T}{\pi d^3} \tag{19.60}$$

Real materials are not *perfectly* rigid but have instead a *modulus of rigidity*, G, which expresses the finite ratio between τ and *shear strain*, γ. The maximum strain in a solid round shaft therefore also exists at its surface and can be found from

$$\gamma_m = \frac{\tau_m}{G} = \frac{16T}{\pi d^3 G} \tag{19.61}$$

Figure 19.45(b) shows the manifestation of shear strain as an angular displacement between axially separated cross sections. Over the length L, the solid round shaft shown will be *twisted* by the torque through an angle ϕ found from

$$\phi = \frac{32LT}{\pi d^4 G} \tag{19.62}$$

Work, Energy, and Power

If during the time of application of a torque, T, the body rotates through some angle θ, mechanical work

$$W = T\theta \tag{19.63}$$

is performed. If the torque acts in the same CW or CCW sense as the displacement, the work is said to be done *on* the body, or else it is done *by* the body. Work done *on* the body causes it to accelerate, thereby appearing as an increase in *kinetic energy* (KE = $I\omega^2/2$). Work done *by* the body causes deceleration with a corresponding decrease in kinetic energy. If the body is not accelerating, any work done on it at one location must be done by it at another location. Work and energy are each measured in units called a joule (J). Equation (19.63) shows that 1 J is equivalent to 1 Nm rad, which, since a radian is a dimensionless ratio, ≡ 1 Nm. To avoid confusion with torque, it is preferable to quantify mechanical work in units of mN, or better yet, in J.

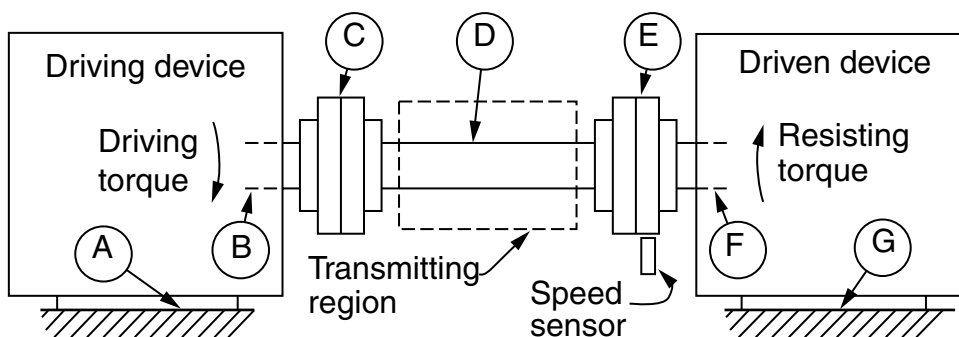



FIGURE 19.46 Schematic arrangement of devices used for the measurement of torque and power.

The *rate* at which work is performed is termed *power*, *P*. If a torque *T* acts over a small interval of time Δt, during which there is an angular displacement $\Delta\theta$, work equal to $T\Delta\theta$ is performed at the rate $T\Delta\theta/\Delta t$. Replacing $\Delta\theta/\Delta t$ by ω, power is found simply as

$$P = T\omega \tag{19.64}$$

The unit of power follows from its definition and is given the special name watt (W). 1 W = 1 J s^{-1} = 1 mN s^{-1}. Historically, power has also been measured in horsepower (Hp), where 1 Hp = 746 W. Rotating bodies effectively transmit power between locations where torques from external sources are applied.

Arrangements of Apparatus for Torque and Power Measurement

Equations (19.58) through (19.64) express the physical bases for torque and power measurement. Figure 19.46 illustrates a generalized measurement arrangement. The actual apparatus used is selected to fulfill the specific measurement purposes. In general, a driving torque originating within a device at one location (B in Fig. 19.46) is resisted by an opposing torque developed by a different device at another location (F). The driving torque (from, e.g., an electric motor, a gasoline engine, a steam turbine, muscular effort, etc.) is coupled through connecting members C, transmitting region D, and additional couplings, E, to the driven device (an electric generator, a pump, a machine tool, mated threaded fasteners, etc.) within which the resisting torque is met at F. The torque at B or F is the quantity to be measured. These torques may be *indirectly* determined from a correlated physical quantity, e.g., an electrical current or fluid pressure associated with the operation of the driving or driven device, or more directly by measuring either the *reaction* torque at A or G, or the *transmitted* torque through D. It follows from the cause-and-effect relationship between torque and rotational motion that most interest in transmitted torque will involve rotating bodies.

To the extent that the frames of the driving and driven devices and their mountings to the "Earth" are *perfectly* rigid, the reaction at A will *at every instant* equal the torque at B, as will the reaction at G equal the torque at F. Under equilibrium conditions, these equalities are independent of the compliance of any member. Also under equilibrium conditions, and except for usually minor *parasitic* torques (due, e.g., to bearing friction and air drag over rapidly moving surfaces), the driving torque at B will equal the resisting torque at F.

Reaction torque at A or G is often determined, using Eq. (19.58), from measurements of the forces acting at known distances fixed by the apparatus. Transmitted torque is determined from measurements, on a suitable member within region D, of τ_m, γ_m, or ϕ and applying Eqs. (19.60)–(19.62) (or analogous expressions for members having other than solid round cross sections [2]). *Calibration*, the measurement of the stress, strain, or twist angle resulting from the application of a *known* torque, makes it unnecessary to know any details about the member within D. When $\alpha \neq 0$, and is measurable, *T* may

also be determined from Eq. (19.59). Requiring only noninvasive, observational measurements, this method is especially useful for determining transitory torques; for example those associated with firing events in multicylinder internal combustion engines [3].

Equations (19.63) and (19.64) are applicable *only* during rotation because, in the absence of motion, no work is done and power transfer is zero. Equation (19.63) can be used to determine *average* torque from calorimetric measurements of the heat generated (equal to the mechanical work W) during a totalized number of revolutions ($\equiv \theta/2\pi$). Equation (19.64) is routinely applied in power measurement, wherein T is determined by methods based on Eqs. (19.58), (19.60), (19.61), or (19.62), and ω is measured by any suitable means [4].

F, T, and ϕ are sometimes measured by simple mechanical methods. For example, a "torque wrench" is often used for the controlled tightening of threaded fasteners. In these devices, torque is indicated by the position of a needle moving over a calibrated scale in response to the elastic deflection of a spring member, in the simplest case, the bending of the wrench handle [5]. More generally, instruments, variously called *sensors* or *transducers*, are used to convert the desired (torque or speed related) quantity into a linearly proportional electrical signal. (Force sensors are also known as *load cells*.) The determination of P most usually requires multiplication of the two signals from separate sensors of T and ω. A transducer, wherein the amplitude of a *single* signal proportional to the power being transmitted along a shaft, has also been described [6].

Torque Transducer Technologies

Various physical interactions serve to convert F, τ, γ, or ϕ into proportional electrical signals. Each requires that some axial portion of the shaft be dedicated to the torque sensing function. Figure 19.47 shows typical features of sensing regions for four sensing technologies in present use.

Surface Strain

Figure 19.47(a) illustrates a sensing region configured to convert surface strain (γ_m) into an electrical signal proportional to the transmitted torque. Surface strain became the key basis for measuring both force and torque following the invention of bonded wire strain gages by E. E. Simmons, Jr. and Arthur C. Ruge in 1938 [7]. A modern strain gage consists simply of an elongated electrical conductor, generally formed in a serpentine pattern in a very thin foil or film, bonded to a thin insulating carrier. The carrier is attached, usually with an adhesive, to the surface of the load carrying member. Strain is sensed as a change in gage resistance. These changes are generally too small to be accurately measured directly and so it is common to employ two to four gages arranged in a Wheatstone bridge circuit. Independence from axial and bending loads as well as from temperature variations are obtained by using a four-gage bridge comprised of two diametrically opposite pairs of matched strain gages, each aligned along a *principal strain* direction. In round shafts (and other shapes used to transmit torque), tensile and compressive principal strains occur at 45° angles to the axis. Limiting strains, as determined from Eq. (19.61) (with τ_m equal to the shear proportional limit of the shaft material), rarely exceed a few parts in 10^3. Typical practice is to increase the compliance of the sensing region (e.g., by reducing its diameter or with hollow or specially shaped sections) in order to attain the limiting strain at the highest value of the torque to be measured. This maximizes the measurement sensitivity.

Twist Angle

If the shaft is *slender* enough (e.g., $L > 5\ d$), ϕ, at limiting values of τ_m for typical shaft materials, can exceed 1°, enough to be resolved with sufficient accuracy for practical torque measurements (ϕ at τ_m can be found by manipulating Eqs. (19.60)–(19.62)). Figure 19.47(b) shows a common arrangement wherein torque is determined from the difference in tooth-space phasing between two identical "toothed" wheels attached at opposite ends of a compliant "torsion bar." The phase displacement of the periodic electrical signals from the two "pickups" is proportional to the peripheral displacement of salient features on the two wheels, and hence to the twist angle of the torsion bar and thus to the torque. These features are chosen to be sensible by any of a variety of noncontacting magnetic, optical, or capacitive techniques.

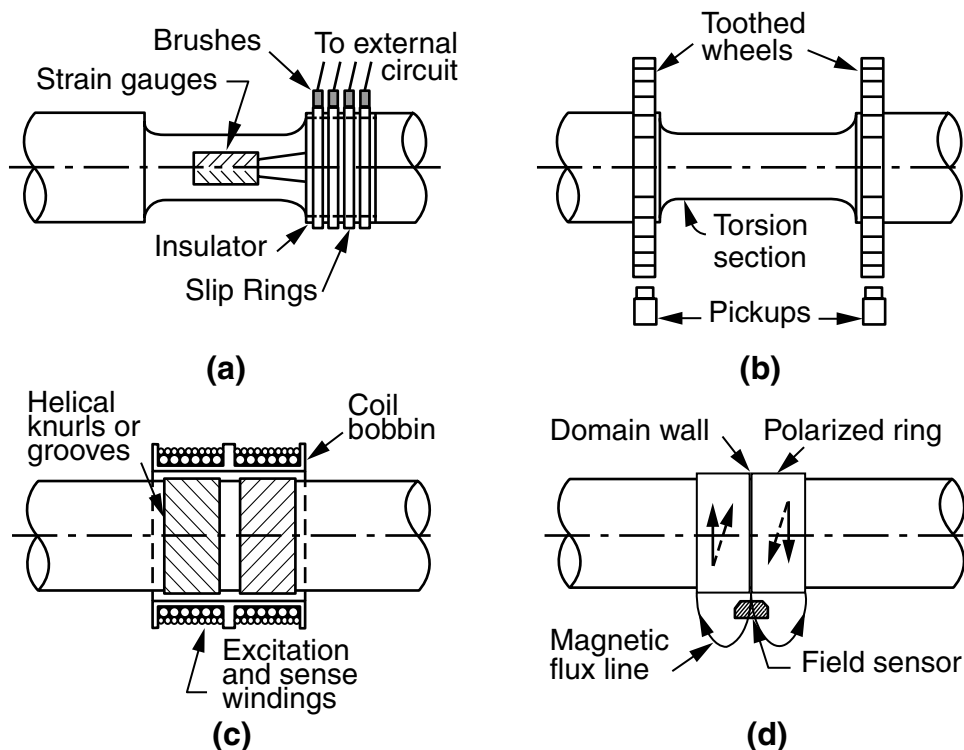



FIGURE 19.47 Four techniques in present use for measuring transmitted torque. (a) Torsional strain in the shaft alters the electrical resistance for four strain gages (two not seen) connected in a Wheatstone bridge circuit. In the embodiment shown, electrical connections are made to the bridge through slip rings and brushes. (b) Twist of the torsion section causes angular displacement of the surface features on the toothed wheels. This creates a phase difference in the signals from the two pickups. (c) The permeabilities of the two grooved regions of the shaft change oppositely with torsional stress. This is sensed as a difference in the output voltages of the two sense windings. (d) Torsional stress causes the initially circumferential magnetizations in the ring (solid arrows) to tilt (dashed arrows). These helical magnetizations cause magnetic poles to appear at the domain wall and ring ends. The resulting magnetic field is sensed by the field sensor.

With more elaborate pickups, the relative angular position of the two wheels appears as the amplitude of a *single* electrical signal, thus providing for the measurement of torque even on a stationary shaft (e.g., [13–15]). In still other constructions, a shaft-mounted variable displacement transformer or a related type of electrical device is used to provide speed independent output signals proportional to ϕ.

Stress

In addition to elastic strain, the stresses by which torque is transmitted are manifested by changes in the magnetic properties of ferromagnetic shaft materials. This "magnetoelastic interaction" [8] provides an inherently noncontacting basis for measuring torque. Two types of magnetoelastic (sometimes called magnetostrictive) torque transducers are in present use: Type 1 derive output signals from torque-induced variations in magnetic circuit permeances; Type 2 create a magnetic field in response to torque. Type 1 transducers typically employ "branch," "cross," or "solenoidal" constructions [9]. In branch and cross designs, torque is detected as an imbalance in the permeabilities along orthogonal 45° helical paths (the principal stress directions) on the shaft surface or on the surface of an *ad hoc* material attached to the shaft. In solenoidal constructions torque is detected by differences in the *axial* permeabilities of two adjacent surface regions, preendowed with symmetrical magnetic "easy" axes (typically along the 45° principal stress directions). While branch and cross-type sensors are readily miniaturized [10], local

variations in magnetic properties of typical shaft surfaces limit their accuracy. Solenoidal designs, illustrated in Fig. 19.47(c), avoid this pitfall by effectively averaging these variations. Type 2 transducers are generally constructed with a ring of magnetoelastically active material rigidly attached to the shaft. The ring is magnetized during manufacture of the transducer, usually with each axial half polarized in an opposite circumferential direction as indicated by the solid arrows in Fig. 19.47(d) [11]. When torque is applied, the magnetizations tilt into helical directions (dashed arrows), causing magnetic poles to develop at the central domain wall and (of opposite polarity) at the ring end faces. Torque is determined from the output signal of one or more magnetic field sensors (e.g., Hall effect, magnetoresistive, or flux gate devices) mounted so as to sense the intensity and polarity of the magnetic field that arises in the space near the ring.

Torque Transducer Construction, Operation, and Application

Although a torque sensing region can be created directly on a desired shaft, it is more usual to install a preassembled *modular* torque transducer into the driveline. Transducers of this type are available with capacities from 0.001 to 200,000 N m. Operating principle descriptions and detailed installation and operating instructions can be found in the catalogs and literature of the various manufacturers [12–20]. Tradenames often identify a specific type of transducers; for example, *Torquemeters* [13] refers to a family of noncontact strain gage models; *Torkducer®* [18] identifies a line of Type 1 magnetoelastic transducers; *Torqstar*™ [12] identifies a line of Type 2 magnetoelastic transducers; *Torquetronic* [16] is a class of transducers using wrap-around twist angle sensors; and *TorXimitor*™ [20] identifies optoelectronic-based, noncontact, strain gage transducers. Many of these devices show generic similarities transcending their specific sensing technology as well as their range. Figure 19.48 illustrates many of these common features.

Mechanical Considerations

Maximum operating speeds vary widely; upper limits depend on the size, operating principle, type of bearings, lubrication, and dynamic balance of the rotating assembly. Ball bearings, lubricated by grease, oil, or oil mist, are typical. Parasitic torques associated with bearing lubricants and seals limit the accuracy of low-end torque measurements. (Minute capacity units have no bearings [15].) Forced lubrication can

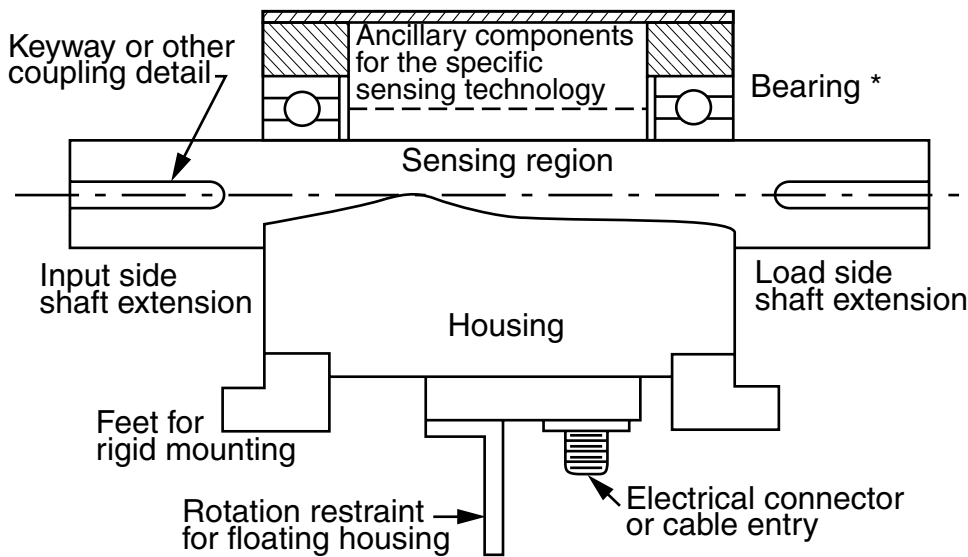



FIGURE 19.48 Modular torque transducer showing generic features and alternative arrangements for free floating or rigid mounting. Bearings* are used only on rotational models. Shaft extensions have keyways or other features to facilitate torque coupling.

allow operation up to 80,000 rpm [16]. High-speed operation requires careful consideration of the effects of centrifugal stresses on the sensed quantity as well as of critical (vibration inducing) speed ranges. Torsional oscillations associated with resonances of the shaft elasticity (characterized by its spring constant) with the rotational inertia of coupled masses can corrupt the measurement, damage the transducer by dynamic excursions above its rated overload torque, and *even be physically dangerous.*

Housings either *float* on the shaft bearings or are *rigidly mounted.* Free floating housings are restrained from rotating by such "soft" means as a cable, spring, or compliant bracket, or by an eccentric external feature simply resting against a fixed surface. In free floating installations, the axes of the driving and driven shafts must be carefully aligned. Torsionally rigid "flexible" couplings at each shaft end are used to accommodate small angular and/or radial misalignments. Alternatively, the use of dual flexible couplings at one end will allow direct coupling of the other end. Rigidly mounted housings are equipped with mounting feet or lugs similar to those found on the frame of electric motors. Free-floating models are sometimes rigidly mounted using adapter plates fastened to the housing. Rigid mountings are preferred when it is difficult or impractical to align the driving and driven shafts, as for example when driving or driven machines are changed often. Rigidly mounted housings *require* the use of dual flexible couplings at *both* shaft ends.

Modular transducers designed for zero or limited rotation applications have no need for bearings. To ensure that *all* of the torque applied at the ends is sensed, it is important in such "reaction"-type torque transducers to limit attachment of the housing to the shaft to only one side of the sensing region. Whether rotating or stationary, the external shaft ends generally include such torque coupling details as flats, keyways, splines, tapers, flanges, male/female squares drives, etc.

Electrical Considerations

By their very nature, transducers require some electrical input power or *excitation.* The "raw" output signal of the actual sensing device also generally requires "conditioning" into a level and format appropriate for display on a digital or analog meter or to meet the input requirements of data acquisition equipment. Excitation and signal conditioning are supplied by electronic circuits designed to match the characteristics of the specific sensing technology. For example, strain gage bridges are typically powered with 10–20 V (DC or AC) and have outputs in the range of 1.5–3.0 mV per volt of excitation at the rated load. Raising these millivolt signals to more usable levels requires amplifiers having gains of 100 or more. With AC excitation, oscillators and demodulators (or rectifiers) are also needed. Circuit elements of these types are normal when inductive elements are used either as a necessary part of the sensor or simply to implement noncontact constructions.

Strain gages, differential transformers, and related sensing technologies require that electrical components be mounted *on* the torqued member. Bringing electrical power to and output signals from these components on rotating shafts require special methods. The most direct and common approach is to use conductive means wherein brushes (typically of silver graphite) bear against (silver) slip rings. Useful life is extended by providing means to lift the brushes off the rotating rings when measurements are not being made. Several "noncontacting" methods are also used. For example, power can be supplied via inductive coupling between stationary and rotating transformer windings [12–15], by the illumination of shaft-mounted photovoltaic cells [20], or even by batteries strapped to the shaft [21] (limited by centrifugal force to relatively low speeds). Output signals are coupled off the shaft through rotary transformers, by frequency-modulated (infrared) LEDs [19,20], or by radio-frequency (FM) telemetry [21]. Where shaft rotation is limited to no more than a few full rotations, as in steering gear, valve actuators or oscillating mechanisms, hard wiring both power and signal circuits is often suitable. Flexible cabling minimizes incidental torques and makes for a long and reliable service life. All such wiring considerations are avoided when noncontact technologies or constructions are used.

Costs and Options

Prices of torque transducers reflect the wide range of available capacities, performance ratings, types, styles, optional features, and accessories. In general, prices of any one type increase with increasing capacity. Reaction types cost about half of similarly rated rotating units. A typical foot-mounted, 565 N m

capacity, strain gage transducer with either slip rings or rotary transformers and integral speed sensor, specified nonlinearity and hysteresis each within ±0.1%, costs about $4000 (1997). Compatible instrumentation prividing transducer excitation, conditioning, and analog output with digital display of torque and speed costs about $2000. A comparable magnetoelastic transducer with ±0.5% accuracy costs about $1300. High-capacity transducers for extreme speed service with appropriate lubrication options can cost more than $50,000. Type 2 magnetoelastic transducers, mass produced for automotive power steering applications, cost approximately $10.

Apparatus for Power Measurement

Rotating machinery exists in specific types without limit and can operate at power levels from fractions of a watt to some tens of megawatts, a range spanning more than 10^8. Apparatus for power measurement exists in a similarly wide range of types and sizes. Mechanical power flows from a *driver* to a *load*. This power can be determined *directly* by application of Eq. (19.64), simply by measuring, in addition to ω, the output torque of the driver or the input torque to the load, whichever is the device under test (DUT). When the DUT is a driver, measurements are usually required over its full service range of speed and torque. The test apparatus therefore must act as a controllable load and be able to *absorb* the delivered power. Similarly, when the DUT is a pump or fan or other type of load, or one whose function is simply to alter speed and torque (e.g., a gear box), the test apparatus must include a *driver* capable of supplying power over the DUT's full rated range of torque and speed. Mechanical power can also be determined *indirectly* by conversion into (or from) another form of energy (e.g., heat or electricity) and measuring the relevant calorimetric or electrical quantities. In view of the wide range of readily available methods and apparatus for accurately measuring both torque and speed, indirect methods need only be considered when special circumstances make direct methods difficult.

Dynamometer is the special name given to the power-measuring apparatus that includes absorbing and/or driving means and wherein torque is determined by the reaction forces on a stationary part (the *stator*). An effective dynamometer is conveniently assembled by mounting the DUT in such a manner as to allow measurement of the reaction torque on its frame. Figure 19.49 shows a device designed to facilitate such measurements. Commercial models (Torque Table® [12]) rated to support DUTs weighing 222–4900 N are available with torque capacities from 1.3 to 226 N m. "Torque tubes" [4] or other DUT mounting arrangements are also used. Other than for possible rotational/elastic resonances, these systems

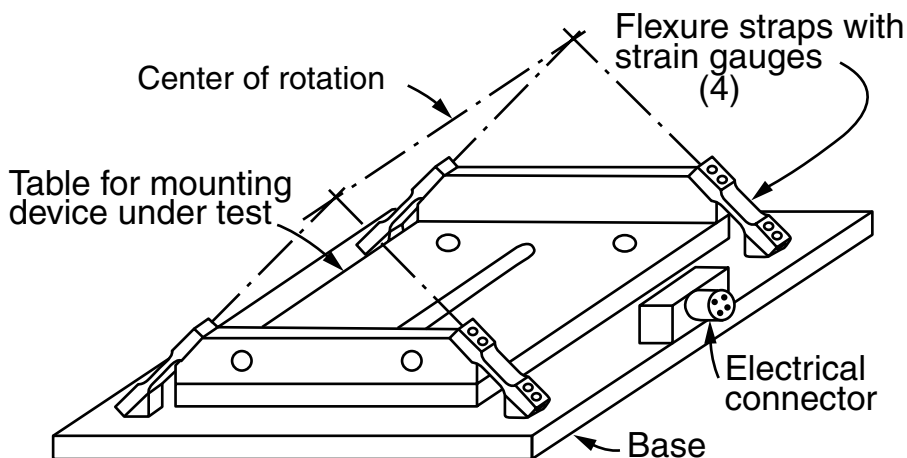



FIGURE 19.49 Support system for measuring the reaction torque of a rotating machine. The axis of the machine must be accurately set on the "center of rotation." The holes and keyway in the table facilitate machine mounting and alignment. Holes in the front upright provide for attaching a lever arm from which calibrating weights may be hung [4,11].

have no speed limitations. More generally, and especially for large machinery, dynamometers include a specialized driving or absorbing machine. Such dynamometers are classified according to their function as *absorbing* or *driving* (sometimes *motoring*). A *universal dynamometer* can function as either a driver or an absorber.

Absorption Dynamometers

Absorption dynamometers, often called *brakes* because their operation depends on the creation of a controllable *drag* torque, convert mechanical work into heat. A drag torque, as distinguished from an active torque, can act only to restrain and not to initiate rotational motion. Temperature rise within a dynamometer is controlled by carrying away the heat energy, usually by transfer to a moving fluid, typically air or water. Drag torque is created by inherently dissipative processes such as: friction between rubbing surfaces, shear or turbulence of viscous liquids, the flow of electrical current, or magnetic hysteresis. Gaspard Riche de Prony (1755–1839), in 1821 [22], invented a highly useful form of a friction brake to meet the needs for testing the steam engines that were then becoming prevalent. Brakes of this type are often used for instructional purposes, for they embody the general principles and major operating considerations for all types of absorption dynamometers. Figure 19.50 shows the basic form and constructional features of a *prony brake.* The power that would normally be delivered by the shaft of the driving engine to the driven load is (for measurement purposes) converted instead into heat via the work done by the frictional forces between the friction blocks and the flywheel rim. Adjusting the tightness of the clamping bolts varies the frictional drag torque as required. Heat is removed from the inside surface of the rim by arrangements (not shown) utilizing either a continuous flow or evaporation of water. There is no need to know the magnitude of the frictional forces nor even the radius of the flywheel (facts recognized by Prony), because, while the drag torque tends to rotate the clamped-on apparatus, it is held stationary by the equal but opposite reaction torque Fr. F at the end of the torque arm of radius r (a fixed dimension of the apparatus) is monitored by a scale or load cell. The power is found from Eqs. (19.58) and (19.64) as $P = Fr\omega = Fr2\pi N/60$ where N is in rpm.

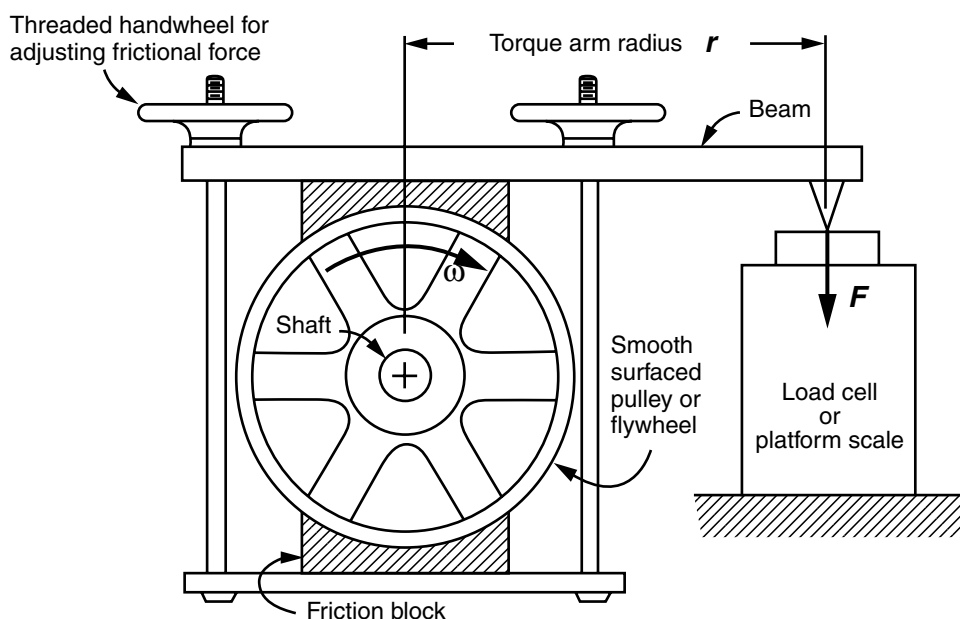



FIGURE 19.50 A classical prony brake. This brake embodies the defining features of all absorbing dynamometers: conversion of mechanical work into heat and determination of power from measured values of reaction torque and rotational velocity.

Uneven retarding forces associated with fluctuating coefficients of friction generally make rubbing friction a poor way to generate drag torque. Nevertheless, because they can be easily constructed, *ad hoc* variations of prony brakes, often using only bare ropes or wooden cleats connected by ropes or straps, find use in the laboratory or wherever undemanding or infrequent power measurements are to be made. More sophisticated prony brake constructions are used in standalone dynamometers with self-contained cooling water tanks in sizes up to 746 kW (1000 Hp) for operation up to 3600 rpm with torques to 5400 N m [23]. Available in stationary and mobile models, they find use in testing large electric motors as well as engines and transmissions on agricultural vehicles. Prony brakes allow full drag torque to be imposed down to zero speed.

William Froude (1810–1879) [24] invented a *water brake* (1877) that does not depend on rubbing friction. Drag torque within a *Froude brake* is developed between the rotor and the stator by the momentum imparted by the rotor to water contained within the brake casing. Rotor rotation forces the water to circulate between cup-like pockets cast into facing surfaces of both rotor and stator. The rotor is supported in the stator by bearings that also fix its axial position. Labyrinth-type seals prevent water leakage while minimizing frictional drag and wear. The stator casing is supported in the dynamometer frame in cradle fashion by *trunnion* bearings. The torque that prevents rotation of the stator is measured by reaction forces in much the same manner as with the prony brake. Drag torque is adjusted by a valve, controlling either the back pressure in the water outlet piping [25] or the inlet flow rate [26] or sometimes (to allow very rapid torque changes) with two valves controlling both [27]. In any case, the absorbed energy is carried away by the continuous water flow. Other types of cradle-mounted water brakes, while externally similar, have substantially different internal constructions and depend on other principles for developing the drag torque (e.g., smooth rotors develop viscous drag by shearing and turbulence). Nevertheless, all *hydraulic dynamometers* purposefully function as *inefficient* centrifugal pumps. Regardless of internal design and valve settings, maximum drag torque is low at low speeds (zero at standstill) but can rise rapidly, typically varying with the square of rotational speed. The irreducible presence of some water, as well as windage, places a speed-dependent lower limit on the *controllable* drag torque. In any one design, wear and vibration caused by cavitation place upper limits on the speed and power level. Hydraulic dynamometers are available in a wide range of capacities between 300 and 25,000 kW, with some portable units having capacities as low as 75 kW [26]. The largest ever built [27], absorbing up to about 75,000 kW (100,000 Hp), has been used to test propulsion systems for nuclear submarines. Maximum speeds match the operating speeds of the prime movers that they are built to test and therefore generally decrease with increasing capacity. High-speed gas turbine and aerospace engine test equipment can operate as high as 30,000 rpm [25].

In 1855, Jean B. L. Foucault (1819–1868) [22] demonstrated the conversion of mechanical work into heat by rotating a copper disk between the poles of an electromagnet. This simple means of developing drag torque, based on *eddy currents*, has, since circa 1935, been widely exploited in dynamometers. Figure 19.51 shows the essential features of this type of brake. Rotation of a toothed or spoked steel rotor through a spatially uniform magnetic field, created by direct current through coils in the stator, induces locally circulating (eddy) currents in electrically conductive (copper) portions of the stator. Electromagnetic forces between the rotor, which is magnetized by the uniform field, and the field arising from the eddy currents, create the drag torque. This torque, and hence the mechanical input power, are controlled by adjusting the *excitation* current in the stator coils. Electrical input power is less than 1% of the rated capacity. The dynamometer is effectively an internally short-circuited generator because the power associated with the resistive losses from the generated eddy currents is dissipated *within* the machine. Being heated by the flow of these currents, the stator must be cooled, sometimes (in smaller capacity machines) by air supplied by blowers [23], but more often by the continuous flow of water [25,27,28]. In *dry gap* eddy current brakes (the type shown in Figure 19.51), water flow is limited to passages within the stator. Larger machines are often of the *water in gap* type, wherein water also circulates around the rotor [28]. Water in contact with the moving rotor effectively acts as in a water brake, adding a nonelectromagnetic component to the total drag torque, thereby placing a lower limit to the controllable torque. Windage limits the minimum value of controllable torque in dry gap types. Since drag torque is developed

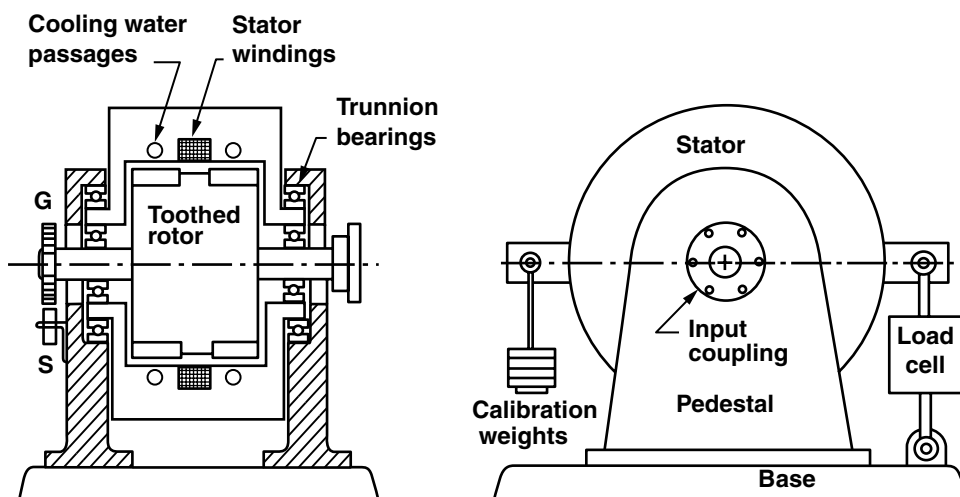



FIGURE 19.51 Cross-section (left) and front view (right) of an eddy current dynamometer. G is a gear wheel and S is a speed sensor. Hoses carrying cooling water and cable carrying electrical power to the stator are not shown.

by the motion of the rotor, it is zero at standstill for any value of excitation current. Initially rising rapidly, approximately linearly, with speed, torque eventually approaches a current limited saturation value. As in other cradled machines, the torque required to prevent rotation of the stator is measured by the reaction force acting at a fixed known distance from the rotation axis. Standard model eddy current brakes have capacities from less than 1 kW [23,27] to more than 2000 kW [27,28], with maximum speeds from 12,000 rpm in the smaller capacity units to 3600 rpm in the largest units. Special units with capacities of 3000 Hp (2238 kW) at speeds to 25,000 rpm have been built [28].

Hysteresis brakes [29] develop drag torque via magnetic attractive/repulsive forces between the magnetic poles established in a reticulated stator structure by a current through the field coil, and those created in a "drag cup" rotor by the stator field gradients. Rotation of the special steel rotor, through the spatial field pattern established by the stator, results in a cyclical reversal of the polarity of its local magnetizations. The energy associated with these reversals (proportional to the area of the hysteresis loop of the rotor material) is converted into heat within the drag cup. Temperature rise is controlled by forced air cooling from a blower or compressed air source. As with eddy current brakes, the drag torque of these devices is controlled by the excitation current. In contrast with eddy current brakes, rated drag torque is available down to zero speed. (Eddy current effects typically add only 1% to the drag torque for each 1000 rpm). As a result of their smooth surfaced rotating parts, hysteresis brakes exhibit low parasitic torques and hence cover a dynamic range as high as 200 to 1. Standard models are available having continuous power capacities up to 6 kW (12 kW with two brakes in tandem cooled by two blowers). Intermittent capacities per unit (for 5 min or less) are 7 kW. Some low-capacity units are convection cooled; the smallest has a continuous rating of just 7 W (35 W for 5 min). Maximum speeds range from 30,000 rpm for the smallest to 10,000 rpm for the largest units. Torque is measured by a strain gage bridge on a moment arm supporting the machine stator.

Driving and Universal Dynamometers

Electric generators, both AC and DC, offer another means for developing a controllable drag torque and they are readily adapted for dynamometer service by cradle mounting their stator structures. Moreover, electrical machines of these types can also operate in a motoring mode wherein they can deliver controllable *active* torque. When configured to operate selectively in either driving or absorbing modes, the machine serves as a universal dynamometer. With DC machines in the absorbing mode, the generated power is typically dissipated in a convection-cooled resistor bank. Air cooling the machine with blowers is usually adequate, since *most* of the mechanical power input is dissipated externally. Nevertheless, *all*

of the mechanical input power is accounted for by the product of the reaction torque and the rotational speed. In the motoring mode, torque and speed are controlled by adjustment of both field and armature currents. Modern AC machines utilize regenerative input power converters to allow braking power to be returned to the utility power line. In the motoring mode, speed is controlled by high-power, solid-state, adjustable frequency inverters. Internal construction is that of a simple three-phase induction motor, having neither brushes, slip rings, nor commutators. The absence of rotor windings allows for higher speed operation than DC machines. Universal dynamometers are "four-quadrant" machines, a term denoting their ability to produce torque in the same or opposite direction as their rotational velocity. This unique ability allows the effective drag torque to be reduced to zero at any speed. Universal dynamometers [25,28] are available in a relatively limited range of capacities (56–450 kW), with commensurate torque (110–1900 N m) and speed (4500–13,500 rpm) ranges, reflecting their principal application in automotive engine development. Special dynamometers for testing transmissions and other vehicular drive train components insert the DUT between a diesel engine or electric motor prime mover and a hydraulic or eddy current brake [30].

Measurement Accuracy

Accuracy of power measurement (see discussion in [4]) is generally limited by the torque measurement (±0.25% to ±1%) since rotational speed can be measured with almost any desired accuracy. Torque errors can arise from the application of extraneous (i.e., not indicated) torques from hose and cable connections, from windage of external parts, and from miscalibration of the load cell. Undetected friction in the trunnion bearings of cradled dynamometers can compromise the torque measurement accuracy. Ideally, well-lubricated antifriction bearings make no significant contribution to the restraining torque. In practice, however, the unchanging contact region of the balls or other rolling elements on the bearing races makes them prone to brinelling (a form of denting) from forces arising from vibration, unsupported weight of attached devices, or even inadvertently during the alignment of connected machinery. The problem can be alleviated by periodic rotation of the (primarily outer) bearing races. In some bearing-in-bearing constructions, the central races are continuously rotated at low speeds by an electric motor while still others avoid the problem by supporting the stator on hydrostatic oil lift bearings [28].

Costs

The wide range of torque, speed, and power levels, together with the variation in sophistication of associated instrumentation, is reflected in the very wide range of dynamometer prices. Suspension systems of the type illustrated in Fig. 19.49 (for which the user must supply the rotating machine) cost $4000–6000, increasing with capacity [12]. A 100-Hp (74.6 kW) *portable* water brake equipped with a strain gage load cell and a digital readout instrument for torque, speed, and power costs $4500, or $8950 with more sophisticated data acquisition equipment [26]. Stationary (and some *transportable* [23]) hydraulic dynamometers cost from $113/kW in the smaller sizes [25] down to $35/kW for the very largest [27]. Transportation, installation, and instrumentation can add significantly to these costs. Eddy current dynamometers cost from as little as $57/kW to nearly $700/kW, depending on the rated capacity, type of control system, and instrumentation [24,25,28]. Hysteresis brakes with integral speed sensors cost from $3300 to $14,000 according to capacity [29]. Compatible controllers, from manual to fully programmable for PC test control and data acquisition via an IEEE-488 interface, vary in price from $500 to $4200. The flexibility and high performance of AC universal dynamometers is reflected in their comparatively high prices of $670–2200/kW [25,28].

References

1. Pinney, C. P. and Baker, W. E., Velocity Measurement, *The Measurement, Instrumentation and Sensors Handbook,* Webster, J. G., Ed., Boca Raton, FL: CRC Press, 1999.
2. Timoshenko, S. *Strength of Materials,* 3rd ed., New York: Robert E. Kreiger, Part I, 281–290; Part II, 235–250, 1956.

3. Citron, S. J., *On-line engine torque and torque fluctuation measurement for engine control utilizing crankshaft speed fluctuations,* U. S. Patent No. 4,697,561, 1987.
4. Supplement to ASME Performance Test Codes, Measurement of Shaft Power, ANSI/ASME PTC 19.7-1980 (Reaffirmed 1988).
5. See, for example, the catalog of torque wrench products of Consolidated Devices, Inc., 19220 San Jose Ave., City of Industry, CA 91748.
6. Garshelis, I. J., Conto, C. R., and Fiegel, W. S., A single transducer for non-contact measurement of the power, torque and speed of a rotating shaft, SAE Paper No. 950536, 1995.
7. Perry, C. C., and Lissner, H. R., *The Strain Gage Primer,* 2nd ed., New York: McGraw-Hill, 1962, 9. (This book covers all phases of strain gage technology.)
8. Cullity, B. D., *Introduction to Magnetic Materials,* Reading, MA: Addison-Wesley, 1972, Section 8.5, 266–274.
9. Fleming, W. J., Magnetostrictive torque sensors—comparison of branch, cross and solenoidal designs, SAE Paper No. 900264, 1990.
10. Nonomura, Y., Sugiyama, J., Tsukada, K., Takeuchi, M., Itoh, K., and Konomi, T., Measurements of engine torque with the intra-bearing torque sensor, SAE Paper No. 870472, 1987.
11. Garshelis, I. J., *Circularly magnetized non-contact torque sensor and method for measuring torque using same,* U.S. Patent 5,351,555, 1994 and 5,520,059, 1996.
12. Lebow® Products, Siebe, plc., 1728 Maplelawn Road, Troy, MI 48099, Transducer Design Fundamentals/ Product Listings, Load Cell and Torque Sensor Handbook No. 710, 1997, also: Torqstar™ and Torque Table®.
13. S. Himmelstein & Co., 2490 Pembroke, Hoffman Estates, IL 60195, MCRT® Non-Contact Strain Gage Torquemeters and Choosing the Right Torque Sensor.
14. Teledyne Brown Engineering, 513 Mill Street, Marion, MA 02738-0288.
15. Staiger, Mohilo & Co. GmbH, Baumwasenstrasse 5, D-7060 Schorndorf, Germany (In the U.S.: Schlenker Enterprises, Ltd., 5143 Electric Ave., Hillside, IL 60162), Torque Measurement.
16. Torquemeters Ltd., Ravensthorpe, Northampton, NN6 8EH, England (In the U.S.: Torquetronics Inc., P.O. Box 100, Allegheny, NY 14707), Power Measurement.
17. Vibrac Corporation, 16 Columbia Drive, Amherst, NH 03031, Torque Measuring Transducer.
18. GSE, Inc., 23640 Research Drive, Farmington Hills, MI 48335-2621, Torkducer®.
19. Sensor Developments, Inc., P.O. Box 290, Lake Orion, MI 48361-0290, 1996 Catalog.
20. Bently Nevada Corporation, P.O. Box 157, Minden, NV 89423, TorXimitor™.
21. Binsfield Engineering, Inc., 4571 W. MacFarlane, Maple City, MI 49664.
22. Gillispie, C. C. (Ed.), *Dictionary of Scientific Biography,* Vol. XI, New York: Charles Scribner's Sons, 1975.
23. AW Dynamometer, Inc., P.O. Box 428, Colfax, IL 61728, Traction dynamometers: portable and stationary dynamometers for motors, engines, vehicle power take-offs.
24. Roy Porter (Ed.), *The Biographical Dictionary of Scientists,* 2nd ed., New York: Oxford University Press, 1994.
25. Froude-Consine, Inc., 39201 Schoolcraft Rd., Livonia, MI 48150, F Range Hydraulic Dynamometers, AG Range Eddy Current Dynamometers, AC Range Dynamometers.
26. Go-Power Systems, 1419 Upfield Drive, Carrollton, TX 75006, Portable Dynamometer System, Go-Power Portable Dynamometers.
27. Zöllner GmbH, Postfach 6540, D-2300 Kiel 14, Germany (In the U.S. and Canada: Roland Marine, Inc., 90 Broad St., New York, NY 10004), Hydraulic Dynamometers Type P, High Dynamic Hydraulic Dynamometers.
28. Dynamatic Corporation, 3122 14th Ave., Kenosha, WI 53141-1412, Eddy Current Dynamometer—Torque Measuring Equipment, Adjustable Frequency Dynamometer.
29. Magtrol, Inc., 70 Gardenville Parkway, Buffalo, NY 14224-1322, Hysteresis Absorption Dynamometers.
30. Hicklin Engineering, 3001 NW 104th St., Des Moines, IA 50322, Transdyne™ (transmission test systems, brake and towed chassis dynamometers).

19.5 Flow Measurement

Richard Thorn

Introduction

Flow measurement is something that nearly everyone has experienced of. Everyday examples include the metering of household utilities such as water and gas. Similarly flowmeters are used in nearly every sector of industry, from petroleum to food manufacture and processing. It is therefore not surprising that today, the total world flowmeter market is worth over $3000 million and expected to continue growing steadily in the future.

However, what is surprising, given the undoubted importance of flow measurement to the economy, is the accuracy and technology of the most commonly used flowmeters which are poor and relatively old fashioned in comparison to instruments used to measure other measurands such as pressure and temperature. For example, the orifice plate flowmeter, which is still one of the most frequently used flowmeters in the process industry, only has a typical accuracy of ±2% of reading and was first used commercially in the late 1800s. The conservative nature of the flow measurement industry means that traditional techniques such as the orifice plate, Venturimeter, and variable area flowmeter still dominate, while ultrasonic flowmeters which were first demonstrated in the 1950s are still considered to be "new" devices by many users. This article will consider the most commonly used commercially available methods of flow measurement. For recent research developments in flow measurement see [1].

Terminology

The term flow measurement is a general term, and before selecting a flowmeter it is important to be sure what type of flow measurement is actually required. For a fluid flowing through a pipe, flow measurement may mean any one of six different types of measurement.

1. Point velocity measurement—the fluid's velocity at a fixed point across the pipe's cross section (m/s)
2. Mean flow velocity measurement—average fluid velocity across the cross section of the pipe (m/s)
3. Volumetric flowrate measurement—the rate of change in the volume of fluid passing through the pipe with time (m^3/s)
4. Total volume measurement—the total volume of fluid which has passed through the pipe (m^3)
5. Mass flowrate measurement—the rate of change in the mass of the fluid passing through the pipe with time (kg/s)
6. Total mass measurement—the total mass of fluid passing through the pipe (kg/s)

Although the most common type of flow measurement is that of a fluid through a closed conduit or pipe, open channel flow measurements are also regularly needed in applications such as sewage and water treatment. For further information on open channel flow measurement techniques see [2].

Flow Characteristics

The fluid being metered is usually a liquid or gas, and is known as single phase flow. However, there is an increasing need for the flowrate of multiphase mixtures to be measured (see the section titled "Two-Phase Flow").

There are a number of important principles relating to the characteristic of flow in a pipe, which should be understood before a flowmeter can be selected and used with confidence. These are the meaning of Reynolds number, and the importance of the flow's velocity profile.

The Reynolds number Re is the ratio of the inertia forces in the flow ($\rho \bar{v} D$) to the viscous forces in the flow (η), and it can be used to determine whether a fluid flow is laminar or turbulent in nature.

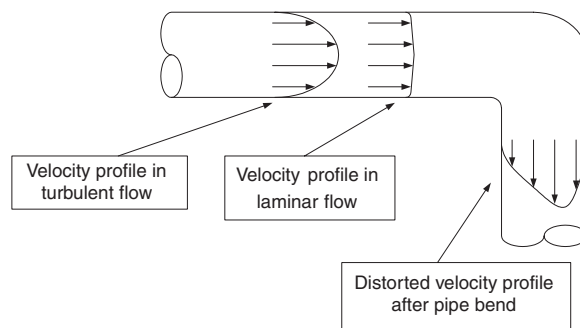



FIGURE 19.52 Flow velocity profiles in laminar and turbulent flow.

Reynolds number can be calculated using

$$\mathrm{Re} = \frac{\rho \bar{v} D}{\eta} \tag{19.65}$$

where ρ is the density of the fluid, $\bar{v}$ is the mean velocity of the fluid, D is the pipe diameter, and η is the dynamic viscosity of the fluid. If Re is less than 2000, viscous forces in the flow dominate and the flow will be laminar. If Re is greater than 4000, inertia forces in the flow dominate and the flow will be turbulent. If Re is between 2000 and 4000 the flow is transitional and either mode can be present. The Reynolds number is mainly calculated using properties of the fluid and does not take into account factors such as pipe roughness, bends, and valves, which also affect the flow characteristic. However, the Reynolds number is a good guide to the type of flow which might be expected in most situations.

The fluid velocity across a pipe's cross section is not constant and depends on the type of flow present (Fig. 19.52). In laminar flow, the velocity at the center of the pipe is twice the average velocity across the pipe cross-section and the flow profile is unaffected by the roughness of the pipe wall. In turbulent flow, pipe wall effects are less and the flow's velocity profile is flatter, with the velocity at the center being about 1.2 times the mean velocity. The exact flow profile in a turbulent flow depends on pipe wall roughness and Reynolds number. In industrial applications laminar flows are rarely encountered unless very viscous fluids are being metered. The pipe Reynolds number should always be calculated since some flowmeters are not suitable for use in both laminar and turbulent flow conditions.

A flow's velocity profile will only be symmetrical at the end of a very long pipe. Bends and obstructions such as valves will cause the profile to become distorted or asymmetric. Since the calibration of many flowmeters is sensitive to the velocity profile of the flow passing through the meter then in order to have confidence in the performance of a flowmeter, the velocity profile of the flow passing through the flowmeter should be stable and known.

Flowmeter Classification

Although there at least 80 different types of flowmeter commercially available, they may be all classified into nine main groups. Table 19.3 gives examples of the main types of flowmeter in each group.

Traditional flow measurement technologies are represented by the differential pressure, variable area, positive displacement, and turbine categories. Newer techniques are represented by the electromagnetic, ultrasonic, oscillatory, and mass categories. Although differential pressure flowmeters are still the most commonly used method of flow measurement, especially in the process industrial sector, in general traditional methods are being increasingly replaced by newer techniques. These techniques are now often preferred because in most cases they do not obstruct the flow, and yet match many of the traditional flowmeters in terms of accuracy and reliability.

TABLE 19.3 Main Categories of Closed Conduit Flowmeter

Type 1—differential pressure flowmeters
Sharp edged orifice plate, chord orifice plate, eccentric orifice plate, Venturi, nozzle, Pitot tube, elbow, wedge, V-cone, Dall tube, Elliot-Nathan flow tube, Epiflo
Type 2—variable area flowmeters
Rotameter, orifice and tapered plug, cylinder and piston, target, variable aperture
Type 3—positive displacement flowmeters
Sliding vane, tri-rotor, bi-rotor, piston, oval gear, nutating-disc, roots, CVM, diaphragm, wet gas
Type 4—turbine flowmeters
Axial turbine, dual-rotor axial turbine, cylindrical rotor, impeller, Pelton wheel, Hoverflo, propeller
Type 5—oscillatory flowmeters
Vortex shedding, swirlmeter, fluidic
Type 6—electromagnetic flowmeters
AC magnetic, pulsed DC magnetic, insertion
Type 7—ultrasonic flowmeters
Doppler, single path transit-time, multi-path transit-time, cross-correlation, drift
Type 8—mass flowmeters
Coriolis, thermal
Type 9—miscellaneous flowmeters
Laser anemometer, hot-wire anemometers, tracer dilution, nuclear magnetic resonance

The following sections will consider the most popular types of flowmeter from each of the eight main categories in Table 19.3. For information on other flowmeters and those in the miscellaneous group see one of the many textbooks on flow measurement such as [3–6].

Differential Pressure Flowmeter

The basic principle of nearly all differential pressure flowmeters is that if a restriction is placed in a pipeline, then the pressure drop across this restriction is related to the volumetric flowrate of fluid flowing through the pipe.

The orifice plate is the simplest and cheapest type of differential pressure flowmeter. It is simply a plate with a hole of specified size and position cut in it, which can then be clamped between flanges in a pipeline (Fig. 19.53). The volumetric flowrate of fluid Q in the pipeline is given by Eq. (19.66):

$$Q = \frac{C}{\sqrt{1-\beta^4}}\varepsilon\frac{\pi}{4}d^2\sqrt{\frac{2(p_1-p_2)}{\rho}} \tag{19.66}$$

where p_1 and p_2 are the pressures on each side of the orifice plate, ρ is the density of the fluid upstream of the orifice plate, d is the diameter of the hole in the orifice plate, and β is the diameter ratio d/D where D is the upstream internal pipe diameter. The two empirically determined correction factors are C the discharge coefficient, and ε the expansibility factor. C is affected by changes in the diameter ratio, Reynolds number, pipe roughness, the sharpness of the leading edge of the orifice, and the points at which the differential pressure across the plate are measured. However, for a fixed geometry it has been shown that C is only dependent on the Reynolds number and so this coefficient can be determined for a particular application. ε is used to account for the compressibility of the fluid being monitored. Both C and ε can be determined from equations and tables in a number of internationally recognized

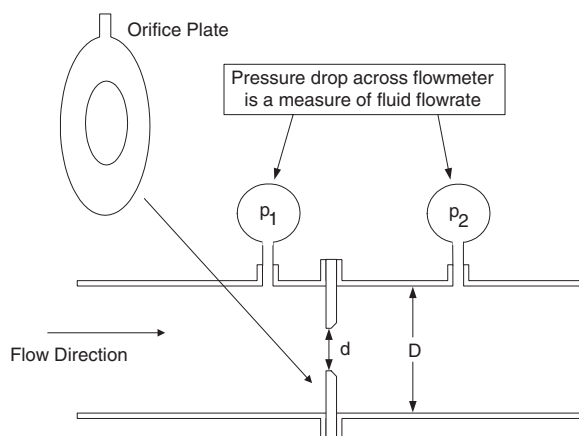



FIGURE 19.53 Flowrate measurement using an orifice plate.

documents known as standards. These standards not only specify C and ε, but also the geometry and installation conditions for the square-edged orifice plate, and two other common types of differential pressure flowmeters, the Venturi tube and nozzle. Installation recommendations are intended to ensure that fully developed turbulent flow conditions exist within the measurement section of the flowmeter. The most commonly used standard in Europe is ISO 5167-1 [7], while in the USA, API 2530 is the most popular [8]. Thus, one of the major reasons for the continued use of the orifice plate flowmeter is that measurement uncertainty (typically ±2% of reading) can be predicted without the need for calibration, as long as it is manufactured and installed in accordance with one of these international standards.

The major disadvantages of the orifice plate are its limited range and sensitivity to changes in the flow's velocity profile. The fact that fluid flow rate is proportional to the square root of the measured differential pressure limits the range of a one plate/one differential pressure transmitter combination to about 3:1. The required diameter ratio (also known as beta ratio) of the plate depends on the maximum flow rate to be measured and the range of the differential pressure transducer available.

Equation (19.66) assumes a fully developed and stable flow velocity profile, and so installation of the device is critical, particularly the need for sufficient straight pipework, upstream and downstream of the meter. Wear of the leading edge of the orifice plate can also severely alter measurement accuracy and so this device is normally only used with clean fluids.

The other two differential pressure flowmeters covered by international standards are the Venturi tube and nozzle. The Venturi tube has a lower permanent pressure less than the orifice plate, and is less sensitive to erosion and upstream disturbances. Major disadvantages are its size and cost. It is more difficult, and therefore more expensive, to manufacture than the orifice plate.

Nozzles have pressure losses similar to orifice plates but because of their smooth design they retain their calibration over a long period. However, these devices are more expensive to manufacture than the orifice plate but cheaper than the Venturi tube. The two most common nozzle designs of nozzle are covered by international Standards, with the ISA-1932 nozzle being preferred in Europe and the ASME long radius nozzle being preferred in the U.S.

There are many other types of differential pressure flowmeter, such as the segmental wedge, V-cone, elbow, and Dall tube. Each of these has advantages over the orifice plate, Venturi tube, and nozzle for specific applications. For example, the segmental wedge can be used with flows having a low Reynolds number, and a Dall tube has a lower permanent pressure loss than a Venturi tube. However, none of these instruments are yet covered by international standards and so calibration is needed to determine their accuracy.

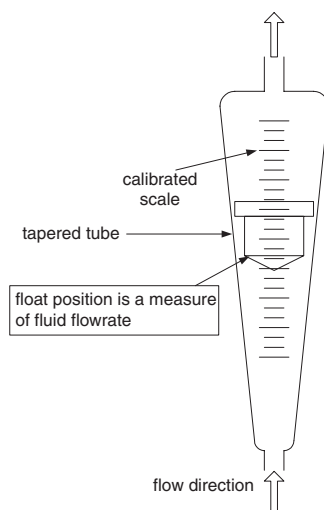



FIGURE 19.54 Tapered tube and float variable area flowmeter.

The Variable Area Flowmeter

Variable area flowmeters are also based on using an obstruction in the flow to create a differential pressure principle, except in this case the differential pressure is constant and the area of the obstruction varies as the flowrate being measured changes. Probably the best known type of variable area flowmeter is the taper tube and float flowmeter, known almost universally as a rotameter (Fig. 19.54). This type of flowmeter consists of a vertical tapered tube into which a float or bob is fitted. The fluid being metered enters the tube at the bottom and forces the float up the tube, which also increases the cross-sectional area available around the float for the fluid to pass through. Increasing the flowrate will move the float further up the tube, and so the position at which the float comes to rest is a direct function of flowrate.

Rotameters are extremely simple and reliable, and have an output which changes linearly with flowrate (unlike differential pressure flowmeters) and a typical range of 10:1 (compared to 3:1 for differential pressure flowmeters). Accuracy is typically ±2% of full scale, but will depend on range and cost of the device. In addition, the flowmeter's calibration is insensitive to changes in the velocity profile of the flow. Since the tube can be made of glass or clear plastic, a visual indication of flowrate is directly available and, of course, the flowmeter requires no external power supply in order to function. As a result such flowmeters are commonly found in many process and laboratory applications where gases or liquids need to be metered. If high temperature, high pressure, or corrosive fluids need to be metered, the rotameter's tube can be made of metal. In such cases a mechanism for detecting and displaying the position of the float is required.

A major limitation of the rotameter is that it can usually only be used vertically and so causes installation difficulties if the pipeline being metered is horizontal. Some manufacturers produce spring loaded rotameters, which can be used in any position; however, in general these have poorer accuracy than standard rotameters. Other limitations are that the calibration of the meter is dependent on the viscosity and density of the fluid being metered, and producing an electrical output signal suitable for transmission requires extra complexity. However, the use of optical or magnetic limit switches to enable the flowmeter to be used in high or low flow alarm applications is common.

The Positive Displacement Flowmeter

Positive displacement flowmeters are based on a simple measurement principle. The flow being measured is "displaced" or moved from the inlet side of the flowmeter to the outlet side using a series of compartments of known volume. The number of compartments of fluid that have been transferred are counted to determine the total volume that has passed through the flowmeter, and if time is also measured then

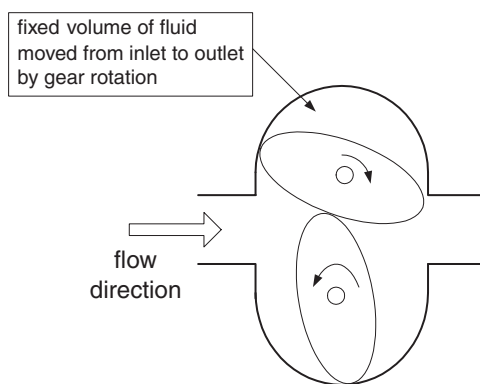



FIGURE 19.55 The oval-gear positive displacement flowmeter.

volumetric flowrate can be measured. There are many designs of positive displacement flowmeters commercially available. For liquids the most common designs are piston, sliding vane, oval-gear, bi-rotor, tri-rotor, and disc types of flowmeter while for gases roots, bellows (or diaphragm), or CVM flowmeters are popular. Despite this wide range of design all are based on the same principle and all are predominantly mechanical devices.

The advantages of positive displacement flowmeters are that they are capable of high accuracy measurement (typically ±0.5% of reading for liquids and ±1% of reading for gases) over a wide range of flowrates. They can be used to meter fluids with a wide range of viscosity and density. In addition, unlike most other flowmeters, they are insensitive to changes in flow velocity profile and so do not require long lengths of straight pipe work before and after the flowmeter.

Figure 19.55 shows the principle of the oval-gear flowmeter and illustrates the limitations of positive displacement flowmeters. They are relatively complex mechanical devices, with moving parts which of course wear with time. Their measurement accuracy depends both on the initial quality of manufacture and a regular maintenance schedule once in use. Fluids being metered should also be free of solid particles so as to reduce wear of the seals and reduce the need for excessive maintenance. Positive displacement flowmeters can also be heavy and expensive for larger pipe sizes and some designs can result in a complete blockage of the pipeline if the flowmeter seizes up.

The Turbine Flowmeter

Like the positive displacement flowmeter, turbine (or vane) flowmeters are mechanical devices capable of achieving high measurement accuracy. The principle of operation of this type of flowmeter is that a multi-bladed rotor is placed in the flow and rotates as fluid passes through it. The rotor's speed of rotation is detected using a sensor (rf, magnetic, and mechanical types being the most common), and is proportional to the velocity of the fluid flowing through the meter. These flowmeters measure the average velocity of fluid in a pipeline, and since the pipe diameter is known, volumetric flowrate can be determined.

Despite the fact that the turbine flowmeter is a mechanical device which may appear to be old fashioned when compared to many other technologies available, it is still one of the most accurate and repeatable flowmeters available today. Measurement accuracy of better than ±0.1% of reading for liquids, and better than ± 0.25% of reading for gases, is possible using this type of flowmeter. For this reason the turbine flowmeter is one of the most commonly used instruments in custody transfer applications. These flowmeters have a linear output and a range of at least 10:1, with 100:1 possible in some applications.

The main limitation of the turbine flowmeter is the fact that key mechanical components such as the rotor bearings will wear with use, and in doing so degrade the instrument's repeatability and alter its calibration. Regular maintenance and recalibration are therefore necessary with this type of flowmeter. Care should also be taken to ensure that the fluid being metered is clean, since solid particles in the flow

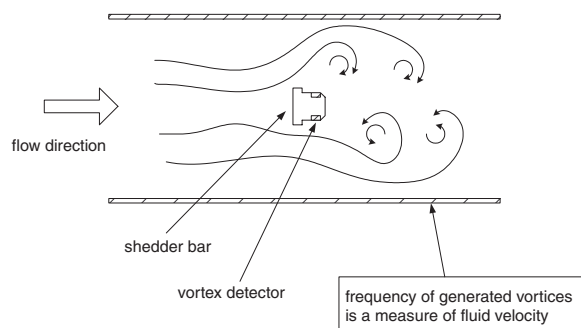



FIGURE 19.56 Principle of the vortex shedding flowmeter.

will cause more rapid bearing wear. The flowmeter's calibration is also sensitive to changes in fluid viscosity and upstream flow velocity profile.

Other types of flowmeters which use the turbine principle include the Pelton wheel and propeller meter, although they have poorer measurement accuracy than axial designs.

The Vortex Shedding Flowmeter

The vortex shedding flowmeter, now more commonly known as the vortex flowmeter, relies on the phenomena of vortex shedding, which was first experimentally studied by Strouhal in 1878. Figure 19.56 shows the principle of the vortex flowmeter. A nonstreamlined obstruction known as a shedder bar or bluff body is placed across the middle of the flow stream. As the fluid stream in the pipe hits this obstacle it must separate to pass around it, with fluid layers nearer the surface of the bar moving slower than those further away. As a result, when these fluid layers recombine after passing the bar, vortices are generated and shed alternately from either side of the shedder bar. The frequency of generated vortices is related to the upstream velocity of the fluid and the width of the shedder bar and is defined by the K factor of the flowmeter. For a given geometry of shedder bar the K factor of a flowmeter is relatively constant over a wide range of pipe Reynolds number, and so in these circumstances the volumetric flowrate of the fluid is linearly related to the vortex shedding frequency.

The frequency of generated vortices is usually detected using sensors integrated into the sides of the shedder bar. Pressure, capacitance, thermal, and ultrasonic are the most common types of sensor used for this purpose.

The vortex flowmeter is capable of accurate measurement of liquid or gas (typically ±1% of reading) over a minimum flow range of 30:1. The flowmeter can also be used over a wide range of fluid temperatures and so is commonly used for metering process fluids at the extreme ends of the temperature range, such as liquid nitrogen and steam. The instrument's calibration is also insensitive to changes in fluid density, and so a meter's calibration holds for any fluid as long as the flowmeter is used within the Reynolds number range specified by the manufacturer. The vortex flowmeter has a simple and reliable construction and so can be used with flows containing small amounts of particles, although more extreme multiphase flows such as slurries will cause rapid wear of the shedder bar and so a change in calibration. The relatively small obstruction that the shedder bar causes results in a permanent pressure loss of about half that of an orifice plate over the same range of flowrate.

The main limitation of the vortex flowmeter is that it can only be used in turbulent flow conditions. It is, therefore, not usually suitable for use in large pipe diameters, or in applications where the flow velocity is low or the fluid viscosity high. Most manufacturers set a minimum Reynolds number of typically 10,000 at which the specified flowmeter performance can be achieved. While many flowmeters will continue operating at Reynolds numbers less than this, the generated vortex stream becomes less stable and so accuracy is reduced. At a Reynolds of less than around 3000, vortices will not be generated

at all and so the flowmeter will stop operating. The vortex flowmeter is also sensitive to changes in upstream flow velocity profile and other disturbances such as swirl, and so should be installed with a sufficient straight length of pipe upstream and downstream of the measurement point. The flowmeter should not be used in applications where pipe vibration or local sources of electrical interference are high, since this will corrupt the vortex signal being detected and possibly give false readings under no-flow conditions.

The Electromagnetic Flowmeter

The operation of the electromagnetic flowmeter is based on Faraday's law of induction, i.e., when a conductor is moving perpendicular to a magnetic field, the voltage induced across the conduction is proportional its velocity. In the case of the electromagnetic flowmeter, the conductor is the fluid being metered, while the induced voltage is measured using electrodes in the pipe wall. Since in most applications the pipe wall of the flowmeter is made from a conductive material such as a stainless steel, an inner nonconducting liner is required to insulate the electrodes and prevent the generated voltage signal being dissipated into the pipe wall. Coils on the outside of the pipe are used to generate a magnetic field across the fluid, with simpler AC coil excitation methods which suffer from zero drift problems being increasingly replaced by pulsed DC excitation techniques which do not.

The electromagnetic flowmeter has a number of advantages over traditional flow measurement techniques, and some characteristics of an ideal flowmeter. The flowmeter has no moving parts, does not obstruct the pipe at all, is available in a very wide range of pipe sizes, and may be used to measure bidirectional flows. A measurement accuracy of typically ±0.5% of reading over a range of at least 10:1 is possible. The flowmeter's accuracy is also unaffected by changes in fluid viscosity and density, and may be used to meter difficult mixtures such as slurries and paper pulp.

The major limitation of the electromagnetic flowmeter is that it can only be used with fluids with a conductivity of typically greater than 5 μS/cm, although special designs are available for use with liquids with conductivities of as low as 0.1 μS/cm. The flowmeter is, therefore, not suitable for use with gases, steam, or nonconducting liquids such as oil. The flowmeter's calibration is also sensitive to changes in flow velocity profile although requiring a shorter straight length of pipe upstream of the meter than the orifice plate or turbine meter. Although electromagnetic flowmeters do not require significant maintenance, care must be taken during operation to ensure that the liner does not become damaged, and that significant deposits do not build-up on the electrodes, since these can cause changes in the calibration or in some cases cause the flowmeter to stop functioning altogether. Even if these effects are minimized, electromagnetic flowmeters will require periodic recalibration using either traditional techniques or an electronic calibrator now available as an accessory from most manufacturers.

The Ultrasonic Flowmeter

The dream of producing a universal non-invasive flowmeter has been the catalyst for the many different ultrasonic flowmeter configurations, which have been investigated over the last 40 years [9]. However, most ultrasonic flowmeters commercially available today can be placed into one of two categories—Doppler and transit-time.

The ultrasonic Doppler flowmeter is based on the Doppler shift principle. Ultrasound at a frequency of typically 1 MHz is transmitted at an angle into the moving fluid being monitored. Some of this energy will be reflected back by acoustic discontinuities such as particles, bubbles, or turbulent eddies. The difference in frequency between the transmitted and received signals (the Doppler frequency shift) is directly proportional to the velocity of the flow.

The ultrasonic transducers, which are used to transmit and receive the ultrasound, are usually located in a single housing that can be fixed onto the outside of the pipe, and so a simple clamp-on flowmeter, which is easy to install and completely noninvasive, is possible.

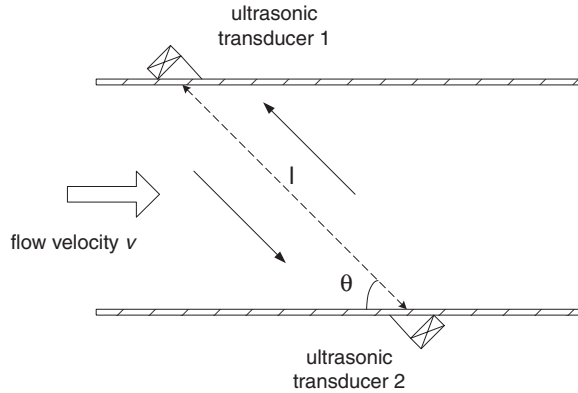



FIGURE 19.57 Principle of the transit-time ultrasonic flowmeter.

The reflected ultrasound does not consist of a single frequency, but a spread of frequencies resulting from reflections from a range of different sized discontinuities, which may also be travelling at different velocities travelling through the detection area. For liquids the frequency of the transmitted ultrasound may lie in the range from 500 kHz up to a few megahertz. At 500 kHz discontinuities must have a diameter of approximately 50 μm in order to reflect ultrasound back to the receiver. Increase in the operating frequency will allow the detection of smaller particles, but at the cost of reducing the penetration of the transmitted signal into the fluid. The flowmeter is also sensitive to changes in flow velocity profile and the spatial distribution of discontinuities in the flow. As a result the accuracy of Doppler flowmeters is poor, typically ±5% of full scale. However, this can be improved by calibrating the flowmeter on-line. Since there is a large acoustic mismatch between steel and air, clamp-on Doppler flowmeters cannot be used for metering gas flows or, of course, totally clean liquids where there are insufficient reflecting particles or bubbles to produce a reliable Doppler signal.

Figure 19.57 illustrates the basic principle of the ultrasonic transit-time flowmeter. Two ultrasonic transducers are mounted on either side of the pipe, so that ultrasound can be transmitted across the fluid flowing in the pipe. The difference in the time it takes for a pulse of ultrasound to travel between transducer 1 and 2 (with the flow) and transducer 2 and 1 (against the flow) is given by

$$\Delta T = \frac{2lv\cos\theta}{c^2 - v^2\cos^2\theta} \tag{19.68}$$

Since the velocity of sound in the fluid c is much greater than the velocity of the fluid v, then

$$\Delta T \approx \frac{2lv\cos\theta}{c^2} \tag{19.69}$$

Therefore, if the velocity of sound in the fluid is constant, then there is a linear relationship between ΔT and v.

Although this method is elegant and straightforward in principle, in practice there are difficulties since ΔT can be small, and the change in ΔT that occurs with changing fluid velocity is even smaller (typically fractions of microsecond per meter). In addition, as Eq. (19.68) shows, if the temperature of the fluid changes then c will change. Measurement of, and correction for, changes in the fluid temperature are usually needed. Transit-time flowmeters, therefore, require the measurement complex signal conditioning and processing.

The flow velocity *v*, which is calculated using Eq. (19.68), is the average velocity along the transmission path between the two transducers, and so the flowmeter's calibration will be very dependent on the flow velocity profile. The problem of flow regime dependency can be significantly reduced by using a configuration of several parallel ultrasonic beams and averaging the measured mean velocity along each beam to calculate the overall fluid velocity. This is analogous to numerical integration, and a wide range of multibeam configurations have been proposed, each with their own advantages.

Single path ultrasonic transit-time flowmeters have a typical accuracy of ±2% of reading over a range of at least 100:1. Although this type of flowmeter can be used with liquids or gases, clamp-on designs can only be used with liquids. Multibeam flowmeters have an improved accuracy, but are more expensive. However, they are finding increased use in high value applications like the custody transfer of natural gas. Unlike most other flowmeters, the cost of transit-time flowmeters does not increase significantly with pipe diameter. Transit-time flowmeters are intended for use with clean fluids, although most can still operate if there are a small amount of impurities present in the flow.

The Coriolis Flowmeter

The Coriolis flowmeter can be used to measure the mass flowrate of a fluid directly. As the name suggests, the principle of operation makes use of the Coriolis effect discovered by Gustave Coriolis in 1835. The fluid being metered passes through a vibrating tube, and as a result of the Coriolis force acting on it, the fluid will accelerate as it moves towards the point of maximum vibration and decelerate as it moves away. This will result in flexure of the tube, the amount of flexure being directly proportional to the mass flowrate of the fluid.

The first commercial Coriolis flowmeter used a U-shaped tube, but now many different configurations exist, including dual loops and straight through designs. Each design has its own advantages, with factors such as accuracy, repeatability, and pressure drop varying from design to design.

Whichever design is used, the Coriolis flowmeter is a high accuracy instrument, which may be used to meter gas or liquid with an accuracy of typically ±0.25% of reading. Measurement range varies with design, but 100:1 is possible for U-tube designs and 50:1 for straight tube designs. Since the flowmeter measures mass directly, changes in density, viscosity, pressure, and temperature do not effect the calibration of the flowmeter. The flowmeter is also not affected by changes in flow velocity profile or other flow disturbances such as swirl. The flowmeter does not obstruct the flow at all, and can be used to meter flow in both directions. However, the pressure drop across U-tube designs can be a limitation with viscous fluids.

The major disadvantage of the Coriolis flowmeter is its cost, which is high in comparison to most other flowmeters. This cost may be justified in applications where the product cost is high, or where mass flowrate of the fluid is required. The cost and weight of the Coriolis flowmeter increase significantly with increasing pipe diameter, and as a result are usually limited to pipe diameters with diameters less than 100 mm.

Unlike most of the flowmeters discussed so far, the Coriolis flowmeter can meter some difficult two-phase flows. For example, reliable measurements of the mass flowrate of liquid/gas mixtures are possible if the gas component is small and well distributed, and therefore the mixture is acting like a pseudo-homogeneous fluid. The percentage of gas that can be tolerated by the flowmeter will depend on the viscosity of the liquid component. The less viscous the liquid the more likely the gas is to separate out and cause problems. Liquid/solids flows (slurries) may also be metered, although the user has to compromise between avoiding particle dropout and avoiding excessive fluid velocities which would result in accelerated wear of the flow tube.

Two-Phase Flow

There is a growing demand in areas such as the petroleum and food industries to be able to measure two-phase flows such as liquid with entrained gas, or liquid with solids. Yet the measurement of such flows nearly always presents difficulties to the process engineer.

For example, an ideal flowmeter would be able to directly measure the volumetric flowrate of a liquid whether it be all liquid or contain a second phase of gas. Unfortunately most of the flowmeters described above can usually only be used to meter two-phase flows when the second component is small. A review of the performance of conventional flowmeters in two-phase flows can be found in [10].

The alternative to direct flow measurement is to use an inferential method. An inferential method for liquid/gas flow would require the velocity of the gas and the liquid phases (v_g and v_l) and cross-sectional fraction of the gas phase (α) to be independently measured in order to calculate the volumetric flowrate of the mixture Q_m:

$$Q_m = v_l A(1 - \alpha) + v_g A \alpha \qquad (19.69)$$

The overall uncertainty of the flowrate measurement would depend on the accuracy with which the independent measurements can be made. The velocity of the liquid and gas phases cannot be assumed to be the same, and the way the gas is distributed in the liquid (the flow regime) will change depending on factors such as gas content, individual phase velocity, and pressure. Even in a simple case such as this, it is clear that multiphase flow measurement is by no means straightforward. For this reason commercial multiphase flowmeters are generally expensive and targeted at very specific applications.

The most common two-phase flows are liquid/gas (e.g., gas in water), liquid/liquid (e.g., water in oil), gas/solid (pneumatically conveyed solids) and liquid/solids or slurries (e.g., coal in oil). Each presents its own measurement problems and it is not feasible to discuss all possible metering combinations for these types of flow in a overview article such as this. For further details of two-phase flow measurement techniques see [11].

Flowmeter Installation

No matter how good a flowmeter is, correct installation is essential if it is to measure with the uncertainty specified by the manufacturer. The calibration of most flowmeters is sensitive to changes in flow velocity profile and in such cases the flowmeter should be installed to ensure that a fully developed and stable velocity profile is present at the measurement point. Manufacturers' data sheets will contain recommendations for the minimum straight length of pipe required before and after a flowmeter to ensure that this is the case. Flow conditioners (or flow straighteners) can be used to correct a distorted velocity profile and remove swirl in applications where long straight lengths are not possible. However, the limitation of all flow conditioners is that they restrict the flow and so produce an unrecoverable pressure drop in the pipeline.

Installation should always ensure that the pipe is completely full at the metering point and that no unwanted second phase is present. In liquid flows entrained gas can be produced from a number of sources, including cavitation and leaking seals. While in a gas flow, an unwanted liquid phase can be produced by processes such as condensation. In most cases flowmeters will produce metering errors if a second phase is present in the flow. While it is possible to remove unwanted phases from the flow upstream of the metering point, it is better to take care with the process, pipework design, and flowmeter installation to ensure that this situation does not occur in the first place.

Flowmeter Selection

There is no such thing as a flowmeter which is equally good for all applications and given the large number of commercial flowmeters and the variety of data sheets available, the choice can sometimes appear bewildering. While at first sight more than one flowmeter may meet a particular application, selecting the most appropriate can be more difficult. In general the best flowmeter will be the one that can meet the performance specification at the lowest total cost (this is a combination of purchase price and cost of maintenance).

There are five factors which can be considered when trying to decide which flowmeter to use. These are the type of fluid to be metered, process conditions, installation conditions, performance requirements, and economic factors. Information required when considering the fluid to be metered are, whether it is a single phase fluid or whether it contains a second component such as gas or solids, the fluid viscosity and density, whether the fluid is corrosive, and if it is a gas whether it is dry or wet. Factors to be considered under process conditions include the pipeline temperature and pressure, and the ambient conditions outside of the pipeline. Installation conditions covers information such as the pipe diameter, the pipe Reynolds number, the orientation of pipework at the measurement point, the length of straight pipework available, whether the flow is pulsating, the need for any flow conditioning, whether an external power source is available, and if the measurement is being made in a hazardous environment. Performance requirements cover the accuracy, repeatability, range, and dynamic response required by the flowmeter. Finally, economic factors cover issues such as the initial cost of the flowmeter, installation cost, maintenance cost, and the type of training required.

Most flow measurement textbooks also contain flowmeter selection charts (for example, [3, 4, 6]), and an international standard is now available on the selection and application of flowmeters [12].

References

1. *Flow Measurement and Instrumentation,* Oxford: Elsevier Science.
2. Grant, D. M., Open channel flow measurement, in D. W. Spitzer (Ed.), *Flow Measurement: Practical Guides for Measurement and Control,* 2nd ed., Research Triangle Park, NC: ISA, 2001.
3. Miller, R. W., *Flow Measurement Engineering Handbook,* 3rd ed., New York: McGraw Hill, 1996.
4. Baker, R. C., *Flow Measurement Handbook,* Cambridge: Cambridge University Press, 2000.
5. Webster, J. G. (Ed.), *Mechanical Variables Measurement: Solid, Fluid, and Thermal,* Boca Raton: CRC Press, 2000.
6. Spitzer, D. W. (Ed.), *Flow Measurement: Practical Guides for Measurement and Control,* 2nd ed., Research Triangle Park, NC: ISA, 2001.
7. International Organisation for Standardization, ISO5167-1, Measurement of Fluid Flow by Means of Pressure Differential Devices—Part 1: Orifice plates, nozzles and Venturi tubes inserted in circular cross-section conduits running full, Geneva, Switzerland, 1991.
8. American Petroleum Institute, API 2530, Manual of Petroleum Measurement Standards Chapter 14—Natural Gas Fluids Measurement, Section 3—Orifice Metering of Natural Gas and Other Related Hydrocarbon Fluids, Washington, 1985.
9. Lynnworth, L. C., *Ultrasonic Measurements for Process Control: Theory, Techniques, Applications,* Boston: Academic Press, 1989.
10. National Engineering Laboratory, UK, *Effects of Two-Phase Flow on Single-Phase Flowmeters,* Flow Measurement Guidance Note No. 3, 1997.
11. Rajan, V. S. V., Ridley, R. K., and Rafa, K. G., Multiphase flow measurement techniques—a review, *Journal of Energy Resource Technology,* 115, 151–161, 1993.
12. British Standards Institution, BS7405, Guide to the Selection and Application of Flowmeters for Measurement of Fluid Flow in Closed Conduits, London, 1991.

19.6 Temperature Measurements

Pamela M. Norris and Bouvard Hosticka

Introduction

Temperature is often cited as the most widely monitored parameter in science and industry, yet the exact definition of temperature is elusive. The simplest definition would relate temperature to the average kinetic energy of the individual molecules that comprise the system. As the temperature increases, the molecular activity also increases, and thus the average kinetic energy increases. This is an adequate definition for

TABLE 19.4 Fixed Points Used in ITS_{90}[*]

Triple point of hydrogen	13.8033 K
Triple point of neon	24.5561 K
Triple point of oxygen	54.3584 K
Triple point of argon	83.8058 K
Triple point of mercury	234.3156 K
Triple point of water	273.16 K
Melting point of gallium	302.9146 K
Freezing point of indium	429.7485 K
Freezing point of tin	505.078 K
Freezing point of zinc	692.677 K
Freezing point of aluminum	933.573 K
Freezing point of silver	1234.93 K
Freezing point of gold	1337.33 K
Freezing point of copper	1357.77 K

[*]Magnum (1990) includes the full definition of these points.

the discussion of temperature measuring techniques presented here. While this definition may help us understand the concept of temperature, it does not help us assign a numerical value to temperature or provide us with a convenient method for measuring temperature. The zeroth law of thermodynamics, formulated in 1931 more than half a century after the first and second laws, lays the foundation for all temperature measurement. It states that if two bodies are in thermal equilibrium with a third body, they are also in thermal equilibrium with each other. By replacing the third body with a thermometer, we can state that two bodies are in thermal equilibrium if both have the same temperature reading even if they are not in contact.

The zeroth law does not enable the assignment of a numerical value for temperature. For that we must refer to a standard scale of temperature. Two absolute temperature scales are defined such that the temperature at zero corresponds to the theoretical state of no molecular movement of the substance. This leads to the Kelvin scale for the SI system and the Rankine scale for the English system. There are other two-point scales derived by identifying two arbitrary defining points for temperature. These are usually defined as the temperature at which a pure substance undergoes a change in phase. Familiar defining points are the freezing and boiling point of water for 0°C and 100°C, respectively. A wide range of such phase changes, many of them triple points where all three phases are in equilibrium, have been accepted as the defining points of the International Practical Temperature Scale of 1990 (ITS_{90}) shown in Table 19.4. These can be used directly as calibration points for temperature monitors as long as the substances are pure and the other conditions, such as pressure, which are included in the defining points are met. Within the ITS_{90} guidelines are standard means of interpolating temperatures between the defined points. For example, platinum resistors are used in the range from 13.8 to 1235 K. The resistance is fitted to the temperature through a higher-order polynomial that may be simplified for more limited ranges between defined temperature points. The difference between a linear interpolation of resistance between the defined points and the higher-order polynomial interpolation never exceeds 2 mK (Magnum and Furukawa, 1990).

Another complication that is encountered in any discussion of temperature measurement is the fact that temperature is an intrinsic rather than an extrinsic property. Thus, temperature can not be added, subtracted, and divided in the same way that measured extrinsic properties such as length or voltage can be manipulated.

Any property that changes predictably in response to temperature can be used in a temperature sensor. The discussion of temperature measuring devices given here subdivides the devices based on the measuring principle. Discussion will begin with a series of thermometers that rely upon the differential expansion coefficients of the materials, be they solid, liquid, or gas. Mercury thermometers, perhaps the most well known and widely used of all temperature measuring devices, belong to this category. We will then move on to devices that rely upon phase change. Next we will discuss electrical temperature sensors and transducers. Included in this category are thermocouples, RTDs, and thermistors, as well as integrated

circuit temperature sensors. The final category of temperature sensors will be noncontact sensors. A separate discussion of temperature measurements on the microscale is provided at the end. Many of the techniques discussed in the microscale section will be derivatives of those introduced earlier in the discussion, but alterations, ranging from minor to quite major, must be made to enable small-scale and/or quick response temperature measurements.

Thermometers That Rely Upon Differential Expansion Coefficients

Thermometers that rely upon differential expansion coefficients are by far the most common and familiar direct reading temperature monitors. These thermometers can be divided into categories depending on the state of the materials used. Each of these deserves a separate discussion.

Gas vs. Solid

The gas bulb thermometer, which is used to determine absolute zero from extrapolation of the change in pressure of a simple gas in a metal sphere with a change in temperature, is an example of a gas vs. solid thermometer. If the metal bulb had the same expansion coefficient as the fill gas, the pressure inside would remain constant and it would not be a thermometer. Instead, the gas follows the ideal gas law, which indicates that, at a constant volume, the pressure is linearly related to the temperature and the vessel containing the gas changes linearly with the volumetric expansion coefficient of the metal making up the bulb. The thermal expansion coefficient of the metal is usually ignored unless very precise predictions of absolute zero are required.

While a large metal sphere with a pressure gage attached is not a very convenient means of measuring temperature, except as a demonstration or research tool, the bulb can be made quite small and connected via a small capillary tube to a remote pressure gage. In this miniaturized configuration the gas bulb thermometer becomes a practical means of measuring temperature. As long as the device operates in the ideal gas region, the pressure gage can be graduated to read temperature directly, since pressure is linearly related to temperature.

Some major limitations on gas bulb thermometers are that the instrument should be calibrated specifically for a particular installation since the length of the heated capillary, as well as the ambient pressure and temperature at the pressure gage, will influence the accuracy of the device. These limitations can be overcome at the expense of complication by using bimetallic elements in the pressure gage to compensate for the temperature at that point or by having a parallel capillary with no bulb follow the main capillary up to the point of measurement and have the parallel capillary equipped with a pressure gage linked to subtract its effects from the main gage. Also, any damage that changes the volume of the bulb, such as a dent, will shift the calibration. This style of instrument should not be confused with vapor pressure thermometers that can take on an identical exterior form, but instead of being filled with an ideal gas, they contain a two-phase fluid and the saturation pressure of the fluid is measured. This type of temperature sensor is discussed in further detail in another section.

Liquid vs. Solid

The common mercury and glass thermometer is an example of a liquid vs. solid temperature sensor. The thermal expansion of liquids, although not as great as gasses, is generally much greater than that of solids and for many applications the expansion coefficient of the glass can be ignored. However, for precision measurements, the expansion of the glass can introduce significant errors. There are two common means of dealing with the glass expansion coefficient. Thermometers intended for reading the temperature of a liquid bath might have a specified submergence depth indicated by a mark on the stem. It is assumed that the rest of the thermometer is at standard lab conditions. This is not always a good assumption, however, and a more precise way of handling the glass expansion coefficient compared to that of mercury is to use a pair of total submergence thermometers. One measures the temperature of the liquid and the other measures the temperature in the immediate vicinity of the exposed stem. A simple stem correction formula supplied by the thermometer manufacturer can then be applied to determine the temperature of the bath.

The ratio of the volume of the bulb to the bore of the capillary determines the resolution of the thermometer. The amount of liquid initially in the thermometer determines its range. The accuracy of the bore and graduation markings determines its precision. The temperature read from a liquid-glass thermometer is only valid if the liquid in the bore is continuous from the bulb to the point of reading. Separations can be eliminated by either contracting the fluid entirely into the bulb or, in some cases, by expanding it into a reservoir at the top of the thermometer. Mercury is useful from near its freezing temperature (about –40°C) to around 500°C. The boiling point of mercury is only 357°C at standard pressure, so mercury thermometers designed for very high temperatures must be pressurized with an inert cover gas when sealed. Alcohol or other liquids can be used in place of mercury, but the accuracy is generally not as great. The temperature range of alcohol extends from about –200°C to +250°C.

An alternative means of using the difference in expansion coefficients of liquids and solids to measure temperature is to use a system filled completely with a liquid and to monitor the change in the volume of the liquid by the position of a bourdon tube or bellows. If the volume-measuring element has a high spring constant, the compressibility of the liquid might have to be considered (Doebelin, 1990). This scheme has the same disadvantages as the gas bulb thermometer discussed above, as well as the same compensation means for overcoming these disadvantages.

Gas, Liquid, and Solid

Some early thermometers consisted of a gas bulb connected to a sealed U-tube containing mercury or another liquid. This, in effect, is really a gas pressure thermometer using a mercury manometer to indicate pressure. To be accurate, the various coefficients of expansion of all three phases must be considered. These instruments are rarely used where accuracy is important. The only example still widely used is a style of minimum-maximum thermometer for ambient air measurements where a small metal fiber is displaced in either leg of the manometer by the mercury. The fiber has enough friction in the tube and is not wetted by the mercury so that it remains free in the glass tube after the mercury shifts. The indicator on the gas bulb side stays at the minimum temperature while the indicator on the other leg stays at the maximum temperature as the mercury recedes. Once the minimum and maximum temperatures are observed, the fibers can be repositioned to the top of the two mercury columns by either centrifugal force (slinging the whole thermometer) or with a magnet if iron fibers are used.

Solid vs. Solid

Bimetallic thermometers consist of two metals with differing temperature expansion coefficients bonded together. As the temperature varies from the temperature at which the metals were initially bonded, the metals expand by differing amounts and the composite experiences a shearing force. The most common means of monitoring the shearing is to allow the metal composite to bend in response to temperature changes. The form of the composite can take on many configurations varying in complexity from a simple leaf fixed at one end with a pointer on the other to a small helix fixed at one end and a turning shaft at the other that is linked to a pointer, possibly through a gear train. The shaft is supported on fine bearings with the pointer as much as a meter away from the bimetallic helix. Stick thermometers with a dial at one end are an example of the latter.

Since the temperature variations produce a force, there must always be some gradated restoring force applied to the bimetallic strip. The most common application is to use the bimetallic strip itself as a restoring spring. The final position of the strip is a balance between the shear imposed by the differing temperature coefficients and the spring constant of the strip. There are instances when bimetallic thermometers are required to actuate a switch. In these cases the load imposed by the switch must be overcome by the shear forces in the strip and the designer must consider it as an external load. The temperature range of bimetallic thermometers is limited by the annealing temperature or phase transformation of the metals. Bimetallics are thus mainly used well below 700°C, and they can be permanently damaged if the metals change their properties or the bonding between the different metals fails. A common pair of metals is a nickel steel, such as Invar with a very low thermal expansion coefficient, bonded to a brass alloy with a high thermal expansion coefficient.

Thermometers That Rely Upon Phase Changes

Phase transitions of pure substances at specified pressures are used in the ITS_{90} to define several of the temperature points. This concept of a phase change being a function of temperature, as well as pressure and the type of material, can be exploited in several forms as a means of determining the temperature of a system by observing either the phase change itself or the conditions at which the two phases are in equilibrium. Several useful applications of this are discussed below.

Liquid to Gas

A common remote thermometer consists of a bulb containing a liquid-gas two-phase fluid connected via a capillary tube to a pressure gage. As long as both phases are present, the pressure read on the gage yields the saturation pressure of the fluid. This arrangement overcomes many of the disadvantages of the gas vs. solid thermometers with the same outward appearance described in section "Gas vs. Solid" above. By monitoring the saturation pressure, the indicated temperature is independent of the temperature of the rest of the system and is insensitive to the actual volume of the bulb and capillary. The fluid is typically an organic solvent such as ethane selected for the particular temperature range desired. To keep the two-phase fluid entirely in the bulb, the pressure can be transmitted through the capillary using a single-phase fluid such as oil. Few fluids have a linear saturation curve. Therefore, most pressure gages have a notably nonlinear scale when graduated into units of temperature. Special compensation springs within the pressure gage can be used to allow for a nearly linear temperature scale, but the extra complication is rarely warranted.

The temperature range of liquid to gas thermometers is limited by the two phases of the fluid and they are typically useful from –40°C to 300°C, although a single instrument rarely will operate over more than about a 150°C span. If the saturation pressure of the fluid is very much greater than 100 kPa, a simple pressure gage referenced to atmospheric pressure can be used. Otherwise best accuracy is obtained using an absolute pressure gage. The volume of the bulb must be large compared to the change in volume of the capillary and bourdon tube in the pressure gage so that both phases are always present in the bulb. The size of the bulb keeps these thermometers from being used for point measurements.

Reversible Phase Change Thermostats

Fixed-point thermostats may be constructed based upon the phase change of a particular sensing element. An example is used in mechanical ice-point references where the sudden expansion of water as it freezes is sensed to cycle the cooling system to maintain a two-phase bath. The actual melting and freezing of the ice maintains the reference temperature. Waxes of various melting points can be used in a similar manner.

Another example is a magnetic switch held closed by a permanent magnet until the Curie temperature is reached at which point the magnet loses its magnetism, or more properly, changes from a ferromagnetic material to a paramagnetic material, and the switch opens. When the material cools, it regains its ferromagnetism, which closes the heater switch. The magnetic material can be selected to have the appropriate Curie temperature.

Fixed Temperature Indicators

Any substance that changes phase at a fixed temperature can be used as a temperature indicator. Numerous examples exist, the most common being a crayon made of a wax with a defined melting point. A mark is made on the object whose temperature needs to be monitored, and if the wax melts, its temperature is higher than the crayon point. These fixed temperature indicators are generally irreversible and can take on many forms in addition to crayons.

A variation that can be either reversible or irreversible is a paint containing suspended solids of the wax or similar material that melts at the desired indication temperature. As long as the particles are solid and scatter light, the paint appears opaque, but when they melt and turn into a liquid with a refractive index close to that of the base paint, it appears clear. These indicators can be made in a series of spots with varying temperature points to help monitor actual temperature attained rather than just

a go, no-go indication. The spots can be made reversible for real-time indications or irreversible to indicate the maximum temperature reached over the monitoring period, although the irreversible ones are much more common than the reversible kind.

Although waxes are widely used, any material that has a distinct phase change at a defined temperature can be used in such a monitor. Imaging techniques other than visual can also be used to determine the phase change. One such application involves using gadolinium or another element that readily absorbs neutrons in a suitable form and observing the melting within the interior of an assembly by means of neutron radioscopy. The temperature range for systems based upon observing melting solids ranges from near ambient to several thousand Kelvin.

Electrical Temperature Sensors and Transducers

A sensor in this context is an element that varies an electrical parameter as a function of temperature. This electrical parameter is then converted to a useful electrical function, such as linear voltage to temperature, with added electronics. The sensor and added electronics make up a transducer. The variation of electrical characteristics with temperature is both a source of measurement possibilities as well as the bane of all electrical measuring systems, since the unwanted change of such things as the gain of an amplifier with temperature causes thermal errors to occur. More effort is expended on eliminating temperature induced electrical variations than is spent exploiting them for temperature measurement.

Thermocouples

There is a relationship between the temperature of a conductor and the kinetic energy of the free electrons. Thus, when a metal is subjected to a temperature gradient, the free electrons will diffuse from the high temperature region to the low temperature region where they have a lower kinetic energy. The electron concentration gradient creates a voltage gradient since the lattice atoms that constitute the positive charges are not free to move. This voltage gradient will oppose the further diffusion of electrons in the wire and a stable equilibrium will be established with no current flow.

The "thermal power" of a material relates the balance of thermal diffusion of the electrons to the electrical conductivity of the metal and is unique for every conductor and usually varies with temperature. The electrical conductivity of the material has a strong influence on the thermal power since it defines the ability of a material to support a voltage gradient. Thus, a SINGLE conductor with its ends at differing temperatures will have a voltage difference between the ends. The trick is to be able to measure the voltage at both ends of the conductor and thus determine the temperature difference between those ends. If we use the same type of wire to measure the voltage across the original wire, the second wire will develop exactly the same voltage difference when its ends are exposed to the same temperatures as the original wire. Therefore, this effect cannot be measured with a pair of similar wires. But because the voltage gradient is a function of the thermal power, which is different for each type of metal, a second conductor of a different type of wire can be used to measure the original voltage gradient. Only a conductor with either no electron mobility or infinite conductivity could be used to measure the absolute voltage gradient associated with the temperature gradient of the original conductor. This is not a practical proposition, so only the difference in the temperature-induced electron gradient between two conductors can ever be measured. This is the basis of thermocouples.

In practical terms, whenever two metals are joined together and the junction is at a different temperature than the free ends of the conductors, the free ends will have a potential difference between them, that is a function of the absolute temperature at the junction and the temperature of the free ends. The relationship between voltage difference and temperature difference will be characteristic of the chosen pair of conductors. Rather than speak of the free ends of the two wires, it is normal to refer to a second junction in the circuit. This is valid and reminds us that there is always a second junction to consider even if the two wires from the thermocouple are attached to a metering circuit. Somewhere within the meter, the circuit is completed and the second junction is formed.

From the previous explanation, all of the classic thermocouple laws can be derived. These laws can be summarized and find application as follows:

Law 1. A third metal introduced in the circuit with both ends of the third metal at an isothermal point does not affect the thermally induced voltage of the original pair.

There are two important implications associated with this law. The first means that the nature of the electrical contact between the wires at the junction is not critical, and that the thermocouple itself can be made up of two wires that are soldered, brazed, welded, or swaged together. In all these cases, a third metal is present at the junction whether it is the filler in soldered and brazed connections, the intermediate alloy produced by welding dissimilar metals, or the metal swage holding the ends of the wire together. This does not mean that there are not other concerns involved with how the junction is formed. It will obviously not do to solder wires together and then use the junction above the melting point of the solder. Most commercially prepared thermocouples are welded for that reason. This law also allows for a metallic item whose temperature is being measured to serve as the actual junction by attaching the thermocouple leads directly to it. This might be done to avoid the time needed to transfer energy between the object and an independent thermocouple assembly. The second implication of this law allows a measuring circuit made of conductors other than those used in the thermocouple to be inserted in the circuit as long as both connections between the measuring circuit and the two thermocouple wires are at the same temperature.

Law 2. The temperatures along the wires do not affect the thermally induced voltage characteristic of the temperature of the two junctions.

This means that the thermocouple leads can be conveniently routed through various temperature regions, and that only the temperatures at the junction and the monitoring location are important in determining the voltage.

Law 3. Each metal has its own voltage gradient for a given temperature gradient independent of the wire used to monitor that voltage.

This means that each type of metal can be calibrated against a standard and that the calibration is valid for each type of thermocouple that can be made from this wire.

Thus far the discussion has been limited to open circuits because this avoids the complications of energy being carried away from a hot junction by the electrons or the I^2R losses in the conductors that both produce heat and reduce the measured voltage. Using high-impedance amplified voltmeters or differential voltmeters having infinite impedance when balanced, allows the practical open circuit voltage to be measured and eliminates these sources of error. However, since the wires used in thermocouples are usually metals with reasonable heat conduction, energy may be inadvertently removed from the measured system by simple heat conduction along the wires.

To actually use a thermocouple the temperature at one junction must be known and some sort of calibration table or polynomial curve fit must be used to convert the measured voltage to temperature at the junction. There are published tables and polynomials for common pairs of metals used in thermocouples referenced to the ice point of water (Croarkin et al., 1993). These are based upon an average alloy but the alloy actually purchased cannot be precisely the same as the one represented by the table. This unavoidable variation of alloy content and application of standard tables is the major source of error in thermocouple readings. Despite the best efforts of the manufacturers, this variation can lead to as much as 2% error in the reported voltage for a given temperature measurement. Individual spools of wire or assembled thermocouples can be calibrated to minimize this source of error.

The temperature of the reference junction must be known to determine the temperature of the measuring junction from the measured voltage. If the temperature of one of the reference junctions is not the same as the reference temperature of the calibration table, the voltage associated with the known temperature of the reference junction can be algebraically added to the measured voltage to determine the voltage that would have been measured if the reference junction were, in fact, at the defined temperature of the table. This is not as difficult as it first appears. If the reference junction is kept in an ice bath, no correction is needed when using the normal calibration tables. To make this easy, there are commercial

ice reference refrigerators specifically designed to be used for thermocouple measurements. There are also electronic sensors that give a voltage or current output that is proportional to its absolute temperature that can be incorporated in a measuring system. These can be used to measure the temperature of the isothermal terminal block of the instrument and then, by either analog circuitry or computer software, the thermocouple voltage associated with that temperature is added to the appropriate measured voltage. This scheme allows direct reading of the voltage associated with the unknown temperature. The requirement that the electronic sensor be at the same temperature as the terminals for the thermocouple cannot be over-emphasized. A multichannel scanner or strip chart recorder may have a terminal board extending over several hundred millimeters in length. If the temperature is electronically measured at one point along this length and a power supply or other heat source in the scanner causes a temperature gradient across the terminal board, serious errors might occur. Care must be taken to assure that what is called an isothermal block is truly isothermal. In the case of the scanner mentioned above, the terminal board and multiplexer had to be removed from its parent instrument and wrapped in insulation before accurate measurements could be obtained.

There are a wide variety of commercial thermocouples in various configurations and materials available depending upon the measuring requirements encountered. As well as thermocouples, a whole industry exists to provide readers, controllers, connectors, wires, and all else that is needed for use. Typical thermocouple readers will have an electronic reference temperature and accommodations for the nonlinear relation between voltage and temperature built in so that it is a simple matter to plug in the matching type of thermocouple and read the temperature. Wires for the standard types are color coded with the outer jacket indicating the wire pair and the color of the individual wires indicating polarity.

It is unfortunate that there is almost a perverse nonuniformity of standards across the world. For example, in the United States (ANSI/MC96-1, 1982) a yellow thermocouple wire sheath indicates a Chromel–Alumel pair with the red lead indicating the negative side when reading elevated temperatures. (In the U.S. system, red is uniformly the negative lead.) Whereas in Japan (JIS-C 1610, 1981), a yellow sheath indicates Iron–Constantan with the red lead representing the positive side.

Oxidation or contamination by unintentional alloying of the wire at high temperature can cause the calibration to shift. The calibration can also shift if the alloy changes along the length of the thermocouple after being subjected to steep temperature gradients at high temperature, which will allow the metals of the alloy to diffuse through one another. For normal uses, however, thermocouples are quite stable, easy to use, and reliable, with types suitable for use from near absolute zero to over 2000°C.

The physical size and thermal mass of the thermocouples define their spatial resolution and time constant, but they can be made with extremely fine wires or films to limit their size down to the micron range at the expense of ruggedness and actual power output. Unlike other electrical thermal sensors, thermocouples can be mounted in direct electrical contact with the measured surface, thereby further improving the time response of the measurement. This concept can be extended by having the actual junction formed by a third metal that is vapor deposited to bridge the insulation between the measuring wires. The junctions then consist of the thin film of metal, which can theoretically reduce the time constant to less than 1 μs (Deobelin, 1990). The main disadvantages of thermocouples are their nonlinear voltage to temperature response, the requirement to know the reference temperature by means other than using a thermocouple, and their relatively low accuracy unless specifically calibrated.

Resistance Temperature Devices (RTDs)

Most materials show a variation in electrical resistance with temperature. For metals, the resistance goes up with temperature in nearly a linear manner. Platinum is the preferred metal for practical resistance temperature measurements and indeed is the specified means of interpolating between the many defined points on the ITS_{90} scale. Metals other than platinum can be used for specific applications. For example, one way of measuring the temperature of the windings in motors or generators is to measure their resistance while operating under load. The copper windings themselves act as RTDs in this case. The discussion that follows is specifically for platinum RTDs, but the concepts apply to all metals.

For precise measurements over a wide temperature range, a higher order polynomial should be applied to determine temperature from resistance. For practical measurements over a narrow range of a few hundred degrees, the resistance can be treated as linear with temperature. The platinum resistors specified by ITS_{90} are not the same as commercially available in probes. ITS_{90} specifies pure, unstrained platinum wire made into sensors, typically with a resistance of 25.5 Ω at 0°C (Mangum and Furukawa, 1990). Most RTD elements use commercially pure platinum with a resistance of 100 Ω or 50 Ω at 0°C, although higher nominal values are sometimes used to minimize the effect of contact and lead resistances in the circuit and lower resistances are used at high temperatures. There are several differing standard platinum curves that reflect varying standard purity platinum alloys. While all alloys are nominally pure platinum, the first-order coefficient for temperatures from 0°C to 100°C of the European standard (DIN 43 760) is 3.85 mΩ/Ω-K, while the American standard is 3.92 mΩ/Ω-K. Besides these two common coefficients, there are several other coefficients listed for pure platinum. When purchasing RTDs and RTD reading equipment, the user should be aware of the standard that applies.

RTD elements can be fabricated as a wire wound onto insulated bobbins or films deposited on insulating substrates. The size of both the support and the wire, as well as any insulating encapsulation, will determine the response time of the element. Films are usually smaller than wire wound sensors and thus have a quicker time response. Due to the need to insulate the RTD from the measured surface, however, and also to avoid straining the element, even small thin film RTDs have time constants that are measured in tenths of seconds. In the special case of using a bare wire to measure gas temperatures, the time constant can be considerably reduced to the tens of microseconds; however, due to the self heating described below, this is more useful as a form of local anemometry rather than temperature measurement.

The measurement of temperature using a RTD is simply a matter of determining its resistance. The relation between resistance and temperature is absolute, so no reference temperature is needed, unlike with thermocouples. However, measuring the resistance of the element is not always simple. There are three conventional techniques, each of which has its own disadvantages, as will be discussed shortly.

Common to all techniques for measuring remote sensors is the complication of the lead wire resistance. As mentioned above, all metals have a variation in resistance with temperature and thus any lead wires will act as RTDs. Thus, if the temperature of the wires between the RTD and the reading mechanism varies, an error will be introduced in the temperature reading unless steps are taken to accommodate the lead temperature effects. For this reason, commercial RTDs are available with two, three, or four wires going to the actual terminals of the resistor, depending on the technique used to handle the lead effects as shown in Fig. 19.58. The arrangement of Fig. 19.58(a) is used where the lead lengths are short,

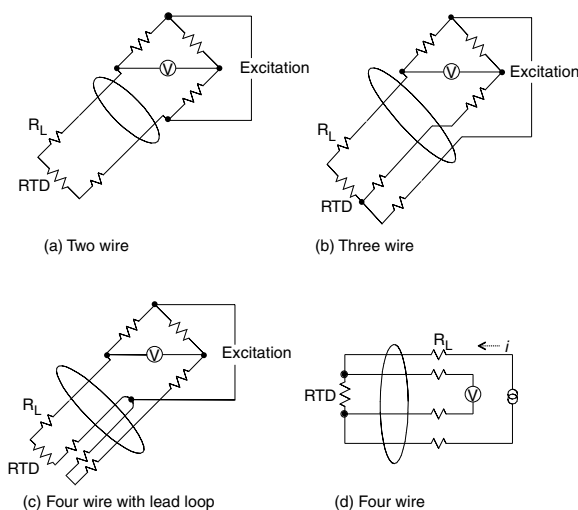



FIGURE 19.58 Various styles of commercial RTD probes and their application in reading circuits.

the resistance of the RTD is high compared to the lead resistance, or where high accuracy is not required. Three wire RTDs, Fig. 19.58(b), allow an equal length of lead wire to be included in each side of the bridge so that changes in lead resistances are felt equally on both sides and the balance of the bridge is not affected. It should be noted that four-wire RTDs of the type shown in Fig. 19.58(d) can be used with readers designed for two-, three-, or four-wire RTDs and unless expense or size is a concern, they are recommended to allow for future upgrades of the reader. A fairly rare alternate configuration of a four wire RTD probe is shown in Fig. 19.58(c) where two of the lead wires form a closed loop. This arrangement usually is needed only when multiple RTDs are in the bridge and should not be confused with the other type of four-wire RTD shown in Fig. 19.58(d), which has two wires going to each side of the resistor.

Another problem common to all reading schemes is that of self-heating of the RTD due to the measuring current through the resistor. This can be minimized by using small currents, having a high heat transfer coefficient between the sensor and the measured process, or by using low duty time pulsed measurements. Although steps can be taken to minimize it, self-heating can never be eliminated. It is such a ubiquitous characteristic of RTDs that it can be exploited as a means of determining the heat transfer coefficient of a system, such as when platinum resistors are used as anemometers.

As indicated above, the measurement of temperature using an RTD simply requires determining its resistance. The three conventional techniques are briefly discussed here.

Fully Balanced Bridges

A primitive but highly accurate means of determining the resistance of an RTD is to use it in one leg of a wheatstone bridge and have a calibrated variable resistor in the opposite leg such that when the bridge is brought to balance, as indicated by a null current on a galvanometer, the resistance of the RTD is the same as the calibrated variable resistor. There are several variations on this theme to eliminate most sources of error by splitting the error-forming device between legs of the bridge. For example, if the RTD is remote from the bridge, a three-wire lead configuration can be used to put an equal length of lead wire in both sides of the bridge with the "corner" of the bridge now defined at the remote RTD location as shown in Fig. 19.58(b). Ultimately the accuracy of the reading is limited by the accuracy of the resistances in the bridge, which can be made quite accurate. Since the variable resistor must be physically adjusted to null the bridge, this technique is not readily adapted to data loggers or temperature displays.

Unbalanced Bridges

If the calibrated variable resistor in the above scheme is replaced by a fixed resistor, the extent of the imbalance of the bridge can be measured with a voltmeter in place of the nulling galvanometer. This technique is fraught with errors since the imbalance of the bridge is not linear with the resistance of the RTD and the reading of the voltmeter is also proportional to the excitation current. Nevertheless, this can be a practical means of measuring temperatures over a narrow range if the values of the resistors in the far legs of the bridge are much higher than the RTD and the opposite leg resistance is close to that of the RTD. As with the other bridge techniques, three lead wires can be arranged in opposite legs of the bridge for remote sensors to compensate for lead resistance changes. If opposite legs of the bridge both consist of RTDs, then the bridge imbalance is proportional to the difference in the temperature of the two RTDs and the only way to keep equal lead lengths in both legs is to use the special four-wire type shown in Fig. 19.58(c).

Direct Voltage vs. Current Measurements

If a constant current is passed through a resistor, the voltage across it is proportional to resistance. This is a simple concept but had to await the advent of accurate constant current sources and high impedance voltage amplifiers to become a practical replacement to the bridge techniques. By using four lead wires, as shown in Fig. 19.58(d), the current circuit can be made independent of the voltage sensing circuit and the lead resistances have no affect on the reading.

Thermistors

Thermistors are bulk semiconductors made from an oxide of nickel, cobalt, manganese, or other metal. The oxide is ground to a fine powder and then sintered to produce the actual thermistor material that is then incorporated into a sensor. Thermistors are resistance temperature sensing devices with several notable differences from RTDs such as their large negative temperature coefficients, and extreme nonlinear response. The resistance of a thermistor is usually so large (several thousand ohms) that lead wire resistance is rarely a concern. Thus, they are inevitably two-wire devices unless multiple thermistors or components are included in the probe. There must be some means of electrical bonding between the wire leads and the thermistor semiconductor. This bonding and the typical epoxy encapsulation places a limit on the maximum usable temperature, even though the thermistor itself is a refractory material.

There are several schemes for dealing with the nonlinearity of thermistors, ranging from applying a correction curve with a computer to having multiple thermistors with differing characteristics complete with nonthermal resistors as a bridge within a single encapsulated probe. For moderate temperatures spanning 200 K, a simple external bridge can be used to linearize the signal.

Although not normally called thermistors, germanium, silicon, and carbon are semiconductors that can also be used to monitor temperatures by measuring their resistance. Germanium is used for very precise measurements at cryogenic temperatures down to less than 1 K. The change in resistance can be very large and very nonlinear, but still very repeatable with a typical unit going from 7000 Ω at 2 K to 6 Ω at 60 K (Doebelin, 1990). Silicon can be used at room temperature and, depending on its doping, can have a very steep temperature curve. It is rarely used as a temperature sensor since other methods work better over its useful range of –200°C to 200°C. Carbon resistors out of a parts drawer found in any laboratory can be used for cryogenic measurements from 1 to 20 K, but they must be individually calibrated.

Integrated Circuit Temperature Sensors

The base-to-emitter voltage drop of a transistor operating at a constant current is a simple function of absolute temperature. Thus, any transistor can be used as a temperature sensor. In reality, this is much more of a problem with building thermally stable electronics than a convenient means of measuring temperature. Integrated circuits are available that monitor the collector current, amplify, and linearize the base-to-emitter voltage to yield an output that is proportional to absolute temperature. Common integrated circuit temperature sensors are available with outputs of 10 mV/K, or 1 μA/K. The temperature range over which they may be used is limited to –50°C to 150°C by the construction techniques of integrated circuits. This makes them very useful for referencing one junction of the thermocouple and most ambient temperature measurements. Although not intrinsically water proof, the ICs are small metal cans or plastic cases resembling signal transistors and can be potted or used in thermowells.

The IC sensors with a voltage output are commonly two terminal devices, with a possible optional lead for trimming the response. When a small current of about 1 mA is allowed to pass through it, it will have a voltage drop directly proportional to the absolute temperature (National, 2000). Even simpler IC transducers are available with separate excitation and signal leads. These are usually calibrated to 10 mV/°F or °C. These have an inherent limitation of not being able to measure below a few degrees above 0°F or 0°C unless both positive and negative power supplies are available.

Voltage output ICs are very convenient where the temperature being monitored is local to the readout and the voltage drop across the lead wires is not a concern, but for remote sensors, which require long lines, current sensors are preferred. Current sensors are also two terminal devices that behave as high impedance current sources so whatever lead resistance present may increase the voltage, but will not affect the current through the sensor (Analog, 1997). Both types can be individually adjusted by trimming resistances on the chip with a laser during manufacture to provide the rated output or they can have an external adjustment lead. Even with trimming and calibration, the accuracy over the entire span from –50°C to +150°C is rarely better than two or three degrees. Several individual ICs may be hooked up to give minimum or average temperature. Voltage ICs are placed in parallel for minimum temperature and in

series for average temperature, while the current devices are connected in series for minimum and parallel for average. In addition to such simple applications of constant current or voltage sources based upon temperature, there are a wide variety of novel circuits to derive almost any function imaginable as a basis of temperature measurement.

Although there is a very small area on the silicon chip of the IC, which is temperature sensitive, it is convenient to regard the entire chip, its case, and the bonded lead wires as the sensor. This increase in thermal mass lowers the time response of the device to several seconds. Self-heating and heat transfer through the leads are also of concern and limit the applicability of these devices in critical measurements.

Noncontact Thermometers

All of the previously discussed temperature monitoring systems implied that the sensor of whatever type is in physical contact with the object being monitored, or in some special cases is the actual object being measured. Often times it is impractical to make this physical connection and noncontact modes of temperature measurement have been developed to overcome this objection. Almost all of these techniques require that the infrared emissions from the surface of the object be measured, but in a few special cases other surface optical properties such as reflectance can be exploited to determine the temperature remotely.

IR Emission Thermometers

Any object above absolute zero emits electromagnetic radiation whose spectrum is related to its surface temperature and surface emissivity. By characterizing the spectrum, the temperature of the object can be determined directly and absolutely. The microwave background of the universe at 3 K, and the temperature-dependent color of stars are extreme examples of this phenomena. Temperature can still be determined from the emitted surface without using a spectrometer. If two bodies are allowed to come into thermal equilibrium with each other and the temperature of one body is known, the temperature of the other is also known. This is the basis of all previously discussed temperature-measuring devices assuming conduction as the principle means of heat transfer. This can be extended to noncontact thermometers since radiation heat transfer is also a valid means of two bodies coming into thermal equilibrium. Many IR thermometers are based upon this phenomenon.

In its simplest form, an IR thermometer would consist of a temperature sensor for monitoring the temperature of an isolated object called the detector, and this detector would only be subject to radiative heat transfer with the surface whose temperature is to be measured. This would work assuming that both the surface and detector behave as black bodies, that there is no heat loss from the detector to the surroundings, and that the field of view of the detector is restricted to the object under measurement and otherwise totally unobstructed. Each one of these assumptions has to be considered when going from the ideal case to a real IR thermometer.

The concept of a black body is an idealization where all radiant energy is completely absorbed by the surface. Under this assumption, the radiant energy is a function only of the temperature of the surface. The only alternatives to being absorbed by the surface are to be reflected by the surface or transmitted through the material. The emissivity, which describes the deviation of a real surface from a black body, is then just one minus its reflectance minus its transmittance. If the emissivity is less than one but independent of wavelength, then it is a gray body. Few real materials are either black bodies or gray bodies, thus emissivity corrections must be made which will often be a function of the temperature being measured. If the surface behaves like a gray body over a limited range of wavelengths, the intensity at a few wavelengths in this range can be measured to estimate the entire spectral shape. It is fortunate that many real objects are close to gray over a narrow range of wavelengths around 750 nm and the spectral shape as a function of temperature for gray objects in the temperature range of 500–3000°C is well enough behaved at these wavelengths that the spectral shape, and thus the temperature, can be measured with just two points. The emissivity of a surface can be determined in conjunction with taking its

temperature by illuminating the surface with an IR laser and measuring the amount of the known laser light that is reflected. The total IR during illumination is the sum of the reflected laser light and the thermally emitted IR. If care is taken that only short pulses or low power laser light is used, to avoid heating the measured surface, this will yield the reflectance at the wavelength of the laser and thus its emissivity which, for opaque surfaces, is one minus reflectance.

Heat transfer to and from the detector by means other than thermal radiation exchange with the monitored surface will require that the temperature of the detector be regarded as only representing the temperature of the monitored surface rather than being the same as the monitored surface. Heat transfer from the detector to the instrument can be via conduction or radiation and is accommodated by calibrating the instrument against a black body of known temperature. When this is done, the temperature of the instrument is usually monitored and circuitry may be set up to have the reference junction for the detector monitor the instrument casing temperature. Alternatively the actual photon flux can be measured by electronic means using photodiodes, photoresistive cells, or other such electronic photon sensors sensitive to IR. When this is done, narrow band filters are often used to limit the response of the detector to a particular wavelength of IR radiation to avoid counting visible photons that may be reflected from the surface or to limit the response to a particular wavelength where atmospheric interferences are minimized.

Optical components are often employed to limit the field of view of the detector so that a defined portion of the surface to be measured is brought to focus on the detector. As long as the entire detector sees the surface of interest, the distance between the detector and the surface is not important, except as it relates to IR absorption by the H_2O, CO_2, or other IR active gasses in the air.

A variation of the IR techniques discussed above is the disappearing filament pyrometer. This device superimposes the image of a tungsten filament whose temperature is a known function of current through the filament onto the view through a telescope. The walls of a furnace or other incandescent surface are observed through the telescope while adjusting the filament current until it just disappears in the background glow. The temperature of the filament then matches the temperature of the incandescent surface and can be determined from the current through the filament. A simple refinement is to put a narrow band red filter in the telescope so that the color is the same for both the target and the filament, and it becomes a single wavelength brightness comparison rather than radiation color comparison. If the emissivity of both the filament and the surface is unity, this can be very accurate. If not equal to one, when a monochromatic filter is used, only the emissivity at that one wavelength needs to be known for accurate temperature determinations.

Microscale Temperature Measurements

As the microelectronics industry surges forward with increasingly higher operating frequencies and increasingly smaller device dimensions, measurement techniques with high spatial and/or temporal resolution are becoming increasingly important. Few techniques are available that can actually measure temperature on a microscale, i.e., sub-micron spatial resolution and/or sub-microsecond temporal resolution. However, many techniques that are being developed concentrate on observing the differential temperature on a microscale. The transient thermoreflectance technique, for example, utilizes a femtosecond pulsed laser to heat and probe the transient reflectance of the sample to enable observation of thermal transport on a sub-picosecond time scale. The technique involves relating the measured reflectivity changes to temperature changes using the material's complex index of refraction (Rosei and Lynch, 1972).

The three most common methods of observing microscale thermal phenomena include thin film thermocouples, thin film microbridges, and optical techniques. Nanometer scale thermocouples are typically used in conjunction with an atomic force microscope (AFM). This technique is nondestructive since the AFM brings the probe into contact with the sample very carefully. Thin film microbridges are patterned metallic thin films, usually thinner than 100 nanometers with a width that depends on the application. This technique relies on the fact that the electrical resistance of the microbridge is a strong function of temperature. The microbridge must be deposited onto the sample surface, therefore the technique is neither noncontact nor nondestructive. Optical techniques typically use a laser as the heating source

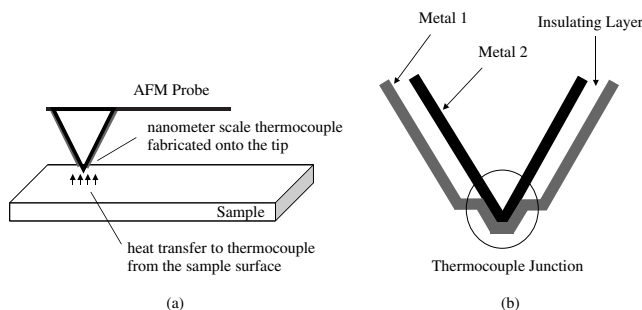



FIGURE 19.59 (a) Diagram showing the use of a scanning thermal microscope probe. (b) Schematic of a nanometer scale thermocouple maufactured onto the tip of a commercially available AFM cantilever.

and/or the thermal probe. The thermal effects can be observed optically in a number of different ways. Thermoreflectance techniques rely on the temperature dependence of reflectance (Paddock and Eesley, 1986), while photothermal techniques monitor the deflection of the probe beam by thermal expansion that results at the surface (Welsh and Ristau, 1995). "Mirage" techniques use the fact that the air just above the surface is also heated, which causes changes in the index of refraction that bends the probe beam by varying amounts depending on the change in temperature (Gonzales et al., 2000).

Scanning Thermal Microscopy (SThM)

The SThM is perhaps the best example of an actual temperature measurement on sub-micron length scales. The nanometer scale thermocouple is comprised of thin metallic films deposited directly onto commercially available AFM probes. Majumdar published a comprehensive review of SThM and includes a description of several methods for manufacturing these nanometer thermocouples (Majumdar, 1999). Figure 19.59(a) shows a diagram of a scanning thermal microscope probe and Fig. 19.59(b) is a schematic of a typical thermocouple junction. There are several factors that affect the spatial resolution of the measurement. These factors include the tip size of the thermocouple which can be on the order of 20 and 50 nm, the mean free path of the energy carrier of the material to be characterized, and the mechanism of heat transfer between the sample and the thermocouple, which is ultimately the limiting factor.

Operation of the AFM cantilever is identical to that of a standard AFM probe. Ideally, the thermocouple would quickly come to thermal equilibrium once in contact with the sample without affecting the temperature of the surface. Practically, a certain amount of thermal energy is transferred between the sample and the thermocouple, which affects the sample temperature, and there is also thermal resistance which delays the measurement and limits the spatial resolution. Once the sample and the thermocouple are brought into contact, there is solid–solid thermal conduction from the sample to the thermocouple. There is also thermal conduction through the gas surrounding the thermocouple tip and through a liquid layer that condenses in the small gap between the tip and the sample. Shi et al. (2000) demonstrated that conduction through this liquid layer dominates the heat transfer under normal atmospheric conditions.

Transient Thermoreflectance (TTR) Technique

The TTR, while not capable of monitoring temperature directly, is an optical technique that enables measurement of temperature changes with sub-picosecond temporal resolution. This technique is fully noncontact and relies on the fact that reflectivity is a function of temperature. The TTR experimental setup (Elsayed-Ali et al., 1991; Paddock and Eesley, 1986; Hostetler et al., 1997) shown in Fig. 19.60 can be employed to monitor the thermoreflectance response of a metallic sample after the absorption of an ultrashort laser pulse. The pulses from a femtosecond laser operating at 76 MHz are separated into two beams, an intense "pump" beam, which is used to heat the film, and a low power "probe" beam, which is used to monitor the reflectivity. The pump beam passes through an acousto-optic modulator that effectively chops

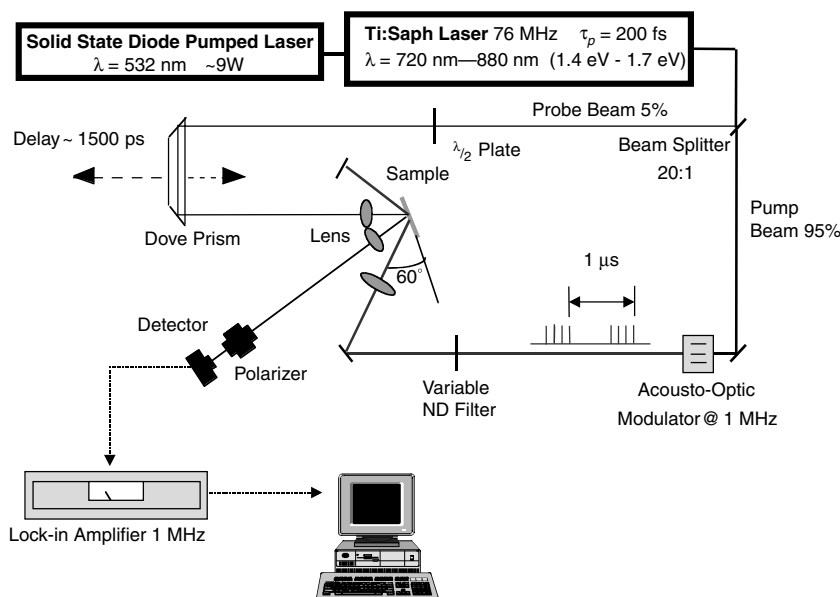



FIGURE 19.60 Experimental setup of the transient thermoreflectance technique.

the beam on and off at a frequency of 1 MHz, resulting in thermomodulation. The probe beam passes through a dovetail prism mounted on a movable stage, which is used to increase the optical path length of the probe beam and hence the time delay between the pump and probe pulses. The reflection of the probe beam, which is centered in the heated area, is monitored by a photodiode and sent to a lock-in amplifier set to the thermomodulation frequency of 1 MHz. This yields the temporal relaxation profile of the sample.

Employing the TTR method as a temperature probe involves relating the measured reflectivity changes to temperature changes using the material's complex index of refraction. In most metals and dielectrics, the complex index of refraction depends weakly on temperature (Price, 1947). In wavelength ranges where the reflection coefficient is large, the reflectivity can be described by the linear sum of a large static contribution and a small temperature-dependent modulated contribution. The corresponding change in reflectivity is $\sim 10^{-5}$/K. The lock-in detection at 1 MHz enables resolution of the small transient signal. By comparing the transient thermal response of a surface to the appropriate heat conduction model, thermophysical properties such as the thermal diffusivity and the thermal boundary resistance can be measured (Hostetler et al., 1997; Hostetler et al., 1998; Smith et al., 2000).

Closing Comments

A wide variety of sensors are available for monitoring the parameter we refer to as temperature. The choice of the appropriate sensor is highly dependent upon the actual physical configuration of the measured material, as well as the required precision, accuracy, and display or processing of the temperature. While thermocouples may be an excellent choice for situations involving electrical logging of a remote process, a gas-bulb thermometer may be adequate and more appropriate for monitoring remote temperatures divorced from electricity. The physical geometry, which often limits access to the area of interest, is another important consideration. It is also important to consider the accuracy requirement, as well as the spatial and temporal resolution desired. This discussion is meant to provide a cursory overview of a wide array of temperature-sensing techniques. There are many excellent, comprehensive references and the designer is referred to these for more details. Temperature measurement often resembles an art rather than a science, with new and creative techniques for monitoring thermal responses in continuous development.

As nanotechnology progresses, many more advances in the area of sub-micron/sub-microsecond temperature measurements will become vital, since many of the traditional means of measuring temperature are not easily adapted to small local temperature measurements.

References

Analog Devices Data Sheet for AD590 Temperature Transducer, 1997.

Croarkin, M.C., et al., 1993, "Temperature-electromotive force reference functions and tables for the letter-designated thermocouple types based on the ITS-90," National Institute of Standards and Technology, Monograph 175.

Doebelin, E.O., *Measurement Systems, Application and Design,* McGraw-Hill, New York, 1990.

Elsayed-Ali, H.E., Juhasz, T., Smith, G.O., and Bron, W.E., 1991, "Femtosecond thermoreflectivity and thermotransmissivity of polycrystalline and single-crystalline gold films," *Phys. Rev. B,* Vol. 43, pp. 4488–4491.

Gonzales, E.J., Bonevich, J.E., Stafford, G.R., White, G., and Josell, D., 2000, "Thermal transport through thin films: mirage technique measurements on aluminum/titanium multilayers," *J. Materials Res.,* Vol. 15, pp. 764–771.

Hostetler, J.L., Smith, A.N., and Norris, P.M., 1998, "Simultaneous measurement of thermophysical and mechanical properties of thin films," *Int. J. Thermophys.,* Vol. 19, pp. 569–577.

Hostetler, J.L., Smith, A.N., and Norris, P.M., 1997, "Thin-film thermal conductivity and thickness measurements using picosecond ultrasonics," *Micro. Thermophys. Eng.,* Vol. 1, pp. 237–244.

Majumdar, A., 1999, "Scanning thermal microscopy," *Ann. Rev. Material Sci.,* Vol. 29, pp. 505–585.

Mangum, B.W. and Furukawa, G.T., 1990, "Guidelines for realizing the international temperature scale of 1990 (ITS-90)," National Institute of Science and Technology, Technical Note 1265.

National Semiconductor Data Sheet for LM135 series Temperature Sensors, DS005698, 2000.

Paddock, C.A. and Eesley, G.L., 1986, "Transient thermoreflectance from thin metal films," *J. Appl. Phys.,* Vol. 60, pp. 285–290.

Price, D.J., 1947, "The temperature variation of the emissivity of metals in the near infrared," *Proc. Phys. Soc.* (London), Vol. 59, pp. 131.

Rosei, R. and Lynch, D.W., 1972, "Thermomodulation spectra of Al, Au, and Cu," *Phys. Rev. B,* Vol. 10, pp. 474–483.

Shi, L., Plyasunov, S., Bachtold, A., McEuen, P.L., and Majumdar, A., 2000, "Scanning thermal microscopy of carbon nanotubes using batch-fabricated probes," *Appl. Phys. Lett.,* Vol. 77, pp. 4295–4297.

Smith, A.N., Hostetler, J.L., and Norris, P.M., 2000, "Thermal boundary resistance measurements using a transient thermoreflectance technique." *Micro. Thermophys. Eng.,* Vol. 4, No. 1, pp. 51–60.

Welsh, E. and Ristau, D., 1995, "Photothermal measurements on optical thin films." *Appl. Opt.,* Vol. 34, pp. 7239–7253.

19.7 Distance Measuring and Proximity Sensors*

Jorge Fernando Figueroa and H. R. (Bart) Everett

Distance Measuring Sensors

Introduction

Range sensors are used to measure the distance from a reference point to an object. A number of technologies have been applied to develop these sensors, the most prominent being light/optics, computer vision, microwave, and ultrasonic. Range sensors may be of contact or noncontact types.

*Significant portions of this chapter were condensed from "Sensors for Mobile Robots", by H. R. Everett, with permission from A. K. Peters, Ltd., Natick, MA.

Noncontact Ranging Sensors

Sensors that measure the actual distance to a target of interest with no direct physical contact are referred to as noncontact ranging sensors. There are at least seven different types of ranging techniques employed in various implementations of such distance measuring devices (Everett et al., 1992):

- Triangulation
- Time of flight (pulsed)
- Phase-shift measurement (CW)
- Frequency modulation (CW)
- Interferometry
- Swept focus
- Return signal intensity

Noncontact ranging sensors can be broadly classified as either *active* (radiating some form of energy into the field of regard) or *passive* (relying on energy emitted by the various objects in the scene under surveillance). The commonly used terms *radar* (radio direction and ranging), *sonar* (sound navigation and ranging), and *lidar* (light direction and ranging) refer to *active* methodologies that can be based on any of several of the above ranging techniques. For example, *radar* is usually implemented using time-of-flight, phase-shift measurement, or frequency modulation. *Sonar* typically is based on time-of-flight ranging, since the speed of sound is slow enough to be easily measured with fairly inexpensive electronics. *Lidar* generally refers to laser-based schemes using time-of-flight or phase-shift measurement.

For any such active (reflective) sensors, effective detection range is dependent not only on emitted power levels, but also the following target characteristics:

- *Cross-sectional area*—determines how much of the emitted energy strikes the target.
- *Reflectivity*—determines how much of the incident energy is reflected versus absorbed or passed through.
- *Directivity*—determines how the reflected energy is redistributed (i.e., scattered versus focused).

Many noncontact sensors operate based on the physics of wave propagation. A wave is emitted at a reference point, and the range is determined by measuring either the propagation time from reference to target, or the decrease of intensity as the wave travels to the target and returns to the reference. Propagation time is measured using time-of-flight or frequency modulation methods.

Ranging by Time-of-Flight (TOF)

Time-of-flight (TOF) is illustrated in Figs. 19.61 and 19.62. A gated wave (a burst of a few cycles) is emitted, bounced back from the target, and detected at the receiver located near the emitter. The emitter and receiver may physically be both one sensor. The receiver may also be mounted on the target. The TOF is the time elapsed from the beginning of the burst to the beginning of the return signal. The distance is defined as $d = c \cdot \text{TOF}/2$ when emitter and receiver are at the same location, or $d = c \cdot \text{TOF}$ when the receiver is attached to the target. The accuracy is usually 1/4 of the wavelength when detecting the return signal, as its magnitude reaches a threshold limit. Gain is automatically increased with distance to maintain accuracy. Accuracy may be improved by detecting the maximum amplitude, as shown in Fig.19.63. This makes detecting the time of arrival of the wave less dependent on the amplitude of the signal. Ultrasonic, RF, or optical energy sources are typically employed; the relevant parameters

FIGURE 19.61 A wave is emitted and bounced from a target object. The distance d is determined from the speed of travel of the wave, c, and the time-of-flight, TOF as $d = (1/2) \cdot c \cdot \text{TOF}$.

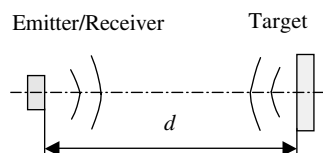
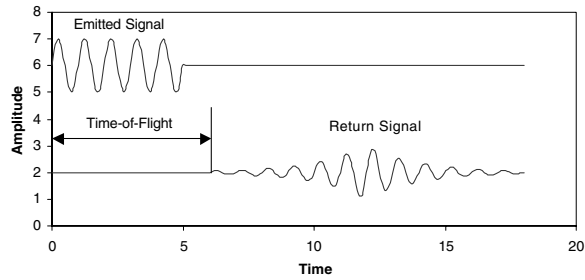



FIGURE 19.62 Definition of Time-of-Flight.

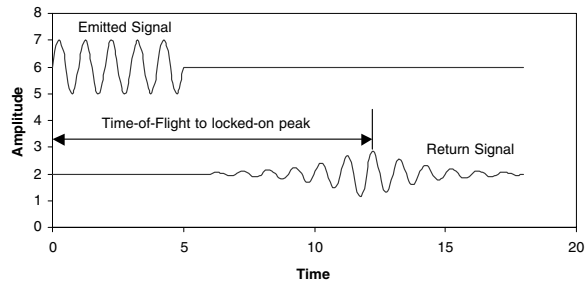



FIGURE 19.63 TOF to the maximum amplitude of the received signal for improved accuracy.

involved in range calculation, therefore, are the speed of sound in air (roughly 0.305 m/ms), and the speed of light (0.305 m/ns).

Potential error sources for TOF systems include the following:

- Variations in the speed of propagation, particularly in the case of acoustical systems
- Uncertainties in determining the exact time of arrival of the reflected pulse (Figueroa & Lamancusa, 1992)
- Inaccuracies in the timing circuitry used to measure the round-trip time of flight
- Interaction of the incident wave with the target surface

Propagation Speed—For most applications, changes in the propagation speed of electromagnetic energy are for the most part inconsequential and can basically be ignored, with the exception of satellite-based position-location systems. This is not the case, however, for acoustically based systems, where the speed of sound is markedly influenced by temperature changes, and to a lesser extent by humidity. (The speed of sound is actually proportional to the square root of temperature in degrees Rankine; an ambient temperature shift of just 30° can cause a 1-ft error at a measured distance of 35 ft.)

Detection Uncertainties—So-called *time-walk errors* are caused by the wide dynamic range in returned signal strength as a result of (1) varying reflectivity of target surfaces, and (2) signal attenuation to the fourth power of distance due to spherical divergence. These differences in returned signal intensity influence the rise time of the detected pulse, and in the case of fixed-threshold detection will cause the less reflective targets to appear further away (Lang et al., 1989). For this reason, *constant fraction timing discriminators* are typically employed to establish the detector threshold at some specified fraction of the peak value of the received pulse (Vuylsteke et al., 1990; Figueroa & Doussis, 1993).

Timing Considerations—The relatively slow speed of sound in air makes TOF ranging a strong contender for low-cost acoustically based systems. Conversely, the propagation speed of electromagnetic energy can place severe requirements on associated control and measurement circuitry in optical or RF implementations. As a result, TOF sensors based on the speed of light require sub-nanosecond timing circuitry to

measure distances with a resolution of about a foot (Koenigsburg, 1982). More specifically, a desired resolution of 1 mm requires a timing accuracy of 3 ps (Vuylsteke et al., 1990). This capability is somewhat expensive to realize and may not be cost effective for certain applications, particularly at close range where high accuracies are required.

Surface Interaction—When light, sound, or radio waves strike an object, any detected echo represents only a small portion of the original signal. The remaining energy reflects in scattered directions and can be absorbed by or pass through the target, depending on surface characteristics and the angle of incidence of the beam. Instances where no return signal is received at all can occur because of specular reflection at the object surface, especially in the ultrasonic region of the energy spectrum. If the transmission source approach angle meets or exceeds a certain critical value, the reflected energy will be deflected outside the sensing envelope of the receiver. Scattered signals can reflect from secondary objects as well, returning to the detector at various times to generate false signals that can yield questionable or otherwise noisy data. To compensate, repetitive measurements are usually averaged to bring the signal-to-noise ratio within acceptable levels, but at the expense of additional time required to determine a single range value.

Ultrasonic TOF Systems

Ultrasonic TOF ranging is today the most common noncontact technique employed, primarily due to the ready availability of low-cost systems and their ease of interface. Over the past few decades, much research has been conducted in investigating applications in mobile robotics for world modeling and collision avoidance, position estimation, and motion detection. Several researchers have assessed the effectiveness of ultrasonic sensors in exterior settings (Pletta et al., 1992; Langer & Thorpe, 1992; Pin & Watanabe, 1993; Hammond, 1994). In the automotive industry, BMW now incorporates four piezoceramic transducers (sealed in a membrane for environmental protection) on both front and rear bumpers in its Park Distance Control system (Siuru, 1994).

The Polaroid ranging module is an active TOF device developed for automatic camera focusing and determines the range to target by measuring elapsed time between transmission of an ultrasonic waveform and the detected echo (Biber et al., 1980). Probably the single most significant sensor development is from the standpoint of its catalytic influence on the robotics research community and industrial applications; this system is the most widely found in the literature (Koenigsburg, 1982; Moravec & Elfes, 1985; Everett, 1985; Kim, 1986; Arkin, 1989; Borenstein & Koren, 1990). Representative of the general characteristics of a number of such ranging devices, the Polaroid unit soared in popularity as a direct consequence of its extremely low cost (Polaroid offers both the transducer and ranging module circuit board for less than $50), made possible by high-volume usage in its original application as a camera auto-focus sensor.

The most basic configuration consists of two fundamental components: (1) the ultrasonic transducer, and (2) the ranging module electronics. A choice of transducer types is now available. In the original instrument-grade electrostatic version (Fig. 19.64), a very thin metalized diaphragm mounted on a machined backplate forms a capacitive transducer (Polaroid, 1981). A smaller diameter electrostatic transducer (*7000-Series*) has also been made available, developed for the Polaroid *Spectra* camera (Polaroid, 1987). A ruggedized piezoelectric (*9000-Series*) *environmental transducer* introduced for applications that may be exposed to rain, heat, cold, salt spray, and vibration is able to meet or exceed guidelines set forth in the SAE J1455 January 1988 specification for heavy-duty trucks. The range of the Polaroid system runs from about 0.3 m (1 ft) out to 10.5 m (35 ft), with a half-power (–3 dB) beam dispersion angle of approximately 12° for the original instrument-grade electrostatic transducer. A typical operating cycle is as follows.

- The control circuitry fires the transducer and waits for an indication that transmission has begun.
- The receiver is blanked for a short period of time to prevent false detection due to residual transmit signal ringing in the transducer.
- The received signals are amplified with increased gain over time to compensate for the decrease in sound intensity with distance.

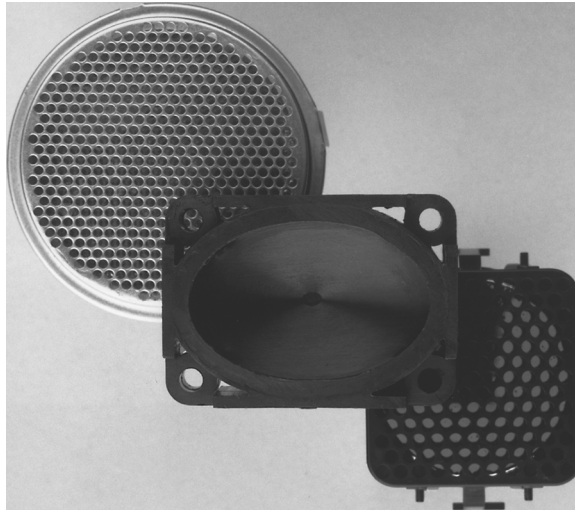



FIGURE 19.64 From left to right: (1) the original instrument grade electrostatic transducer, (2) *9000-Series* environmental transducer, and (3) *7000 Series* electrostatic transducer (courtesy Polaroid Corp.).

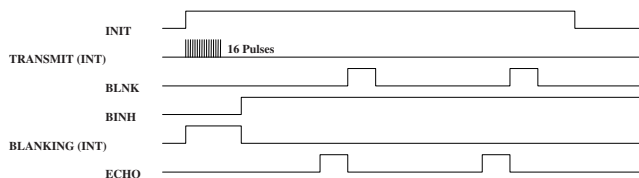



FIGURE 19.65 Timing diagrams for the 6500-*Series Sonar Ranging Module* executing a multiple-echo-mode cycle with blanking input (courtesy Polaroid Corp.).

- Returning echoes that exceed a fixed-threshold value are recorded and the associated distances calculated from elapsed time.

In the *single-echo* mode of operation for the *6500-series* module, the *blank* (BLNK) and *blank-inhibit* (BINH) lines are held low as the *initiate* (INIT) line goes high to trigger the outgoing pulse train. The *internal blanking* (BLANKING) signal automatically goes high for 2.38 ms to prevent transducer ringing from being misinterpreted as a returned echo. Once a valid return is received, the echo (ECHO) output will latch high until reset by a high-to-low transition on INIT. For *multiple-echo* processing, the *blank* (BLNK) input must be toggled high for at least 0.44 ms after detection of the first return signal to reset the *echo* output for the next return, as shown in Fig. 19.65 (Polaroid, 1990).

Laser-Based TOF Systems

Laser-based TOF ranging systems, also known as *laser radar* or *lidar*, first appeared in work performed at the Jet Propulsion Laboratory, Pasadena, CA, in the 1970s (Lewis & Johnson, 1977). Laser energy is emitted in a rapid sequence of short bursts aimed directly at the object being ranged. The TOF of a given pulse reflecting off the object is used to calculate distance to the target based on the speed of light. Accuracies for early sensors of this type could approach a few centimeters over the range of 1–5 m (NASA, 1977; Depkovich & Wolfe, 1984).

Schwartz Electro-Optics, Inc. (SEO), Orlando, FL, produces a number of laser TOF rangefinding systems employing an innovative *time-to-amplitude-conversion* scheme to overcome the sub-nanosecond timing requirements necessitated by the speed of light. As the laser fires, a precision film capacitor begins discharging from a known set point at a constant rate, with the amount of discharge being proportional

FIGURE 19.66 The *LRF-200* series rangefinder (courtesy Schwartz Electro Optics, Inc.).

FIGURE 19.67 The Class 1 (eye-safe) *LD90-3 series* TOF laser rangefinder is a self-contained unit available in several versions with maximum ranges of 150–500 m under average atmospheric conditions (courtesy RIEGL USA).

to the round-trip time-of-flight (Gustavson & Davis, 1992). An analog-to-digital conversion is performed on the sampled capacitor voltage; at the precise instant a return signal is detected, whereupon the resulting digital representation is converted to range and time-walk corrected using a look-up table.

The *LRF-X* series rangefinder shown in Fig. 19.66 features a compact size, high-speed processing, and an ability to acquire range information from most surfaces (i.e., minimum 10% Lambertian reflectivity) out to a maximum of 100 m. The basic system uses a pulsed InGaAs laser diode in conjunction with an avalanche photodiode detector and is available with both analog and digital (RS-232) outputs.

RIEGL Laser Measurement Systems, Horn, Austria, offers a number of commercial products (i.e., laser binoculars, surveying systems, "speed guns," level sensors, profile measurement systems, and tracking laser scanners) employing short-pulse TOF laser ranging. Typical applications include lidar altimeters, vehicle speed measurement for law enforcement, collision avoidance for cranes and vehicles, and level sensing in silos.

The RIEGL *LD90-3 series* laser rangefinder (Fig. 19.67) employs a near-infrared laser diode source and a photodiode detector to perform TOF ranging out to 500 m with diffuse surfaces, and to over 1000 m in the case of cooperative targets. Round-trip propagation time is precisely measured by a quartz-stabilized clock and converted to measured distance by an internal microprocessor, using one of two available algorithms. The *clutter suppression* algorithm incorporates a combination of range measurement averaging and noise rejection techniques to filter out backscatter from airborne particulates, and is, therefore, useful when operating under conditions of poor visibility (Riegel, 1994). The *standard measurement* algorithm, on the other hand, provides rapid range measurements without regard for noise suppression, and can subsequently deliver a higher update rate under more favorable environmental conditions.

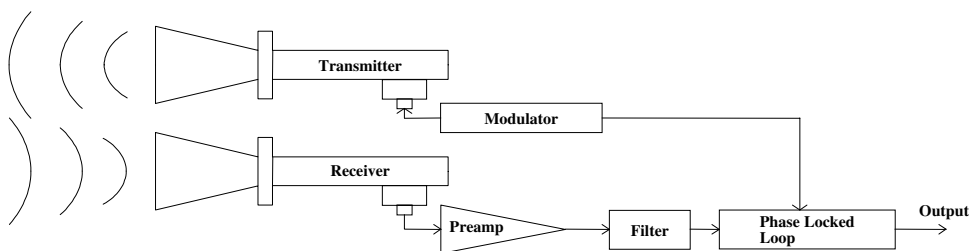



FIGURE 19.68 The microwave sensor, unlike the motion detector, requires a separate transmitter and receiver (adapted from Williams, 1989).

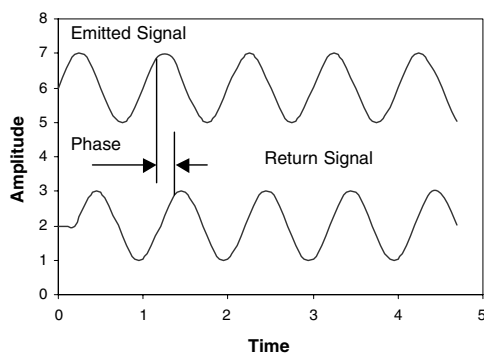



FIGURE 19.69 Range from phase measurement.

Worst-case range measurement accuracy is ±5 cm, with typical values of around ±2 cm. The pulsed near-infrared laser is Class-1 eye-safe under all operating conditions.

Microwave Range Sensors

Microwave technology may be used to measure motion, velocity, range, and direction of motion (Fig. 19.68). The sensors are rugged since they have no moving parts. They can be operated safely in explosive environments, because the level of energy used is very low (no risk for sparks). Their operating temperatures range from −55°C to +125°C. They can work in environments with dust, smoke, poisonous gases, and radioactivity (assuming the components are hardened for radiation). Typically microwave sensors are used to measure ranges from 25 to 45,000 mm, but longer ranges are possible depending on power and object size. The reflected power returning to the receiver decreases as the fourth power of the distance to the object. Typical wavelength used ranges from 1 to 1000 mm.

Time-of-flight is in the order 2 ns per foot of range (reach the target and return). This translates into 10.56 ms per mile of range. Measuring short ranges may pose a problem. For 1 in. resolution, the circuit must resolve 167 ps. An alternate method more suitable to measure short distances is based on a frequency sweep of the signal generator. In this case, the return signal remains at the initial frequency (usually 10.525 GHz), and it is compared with the current frequency changed by a sweep rate. For example, to measure a range of 3 ft, one may sweep at 5 MHz/ms. After 6 ns, the frequency changes by 30 Hz (6 ns × 5 MHz/0.001 s). In this case, 0.0256 mm (0.001 in.) may be resolved easily. When using this method, a signal amplifier that increases gain with frequency is necessary. See section "Frequency Modulation" for more details on frequency modulation methods.

Phase Measurement

Time-of-flight (TOF) is defined as a phase shift between emitted and received signals when the distance is less than one wavelength (Fig. 19.69). Given a phase shift *f*, the distance is calculated as

$d = \phi\lambda/4\pi = \phi c/4\pi f$ if the emitter and receiver are at the same location, or $d = \phi\lambda/2\pi = \phi c/2\pi f$ if the receiver is attached to the target, where c is the speed of travel, ϕ is the measured phase, and f is the modulation frequency.

The phase shift between outgoing and reflected sine waves can be measured by multiplying the two signals together in an electronic mixer, then averaging the product over many modulation cycles (Woodbury et al., 1993). This integrating process can be relatively time consuming, making it difficult to achieve extremely rapid update rates. The result can be expressed mathematically as follows (Woodbury et al., 1993):

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T \sin\left(\frac{2\pi c}{\lambda}t + \frac{4\pi d}{\lambda}\right)\sin\left(\frac{2\pi c}{\lambda}\right)dt \tag{19.70}$$

which reduces to

$$A\cos\left(\frac{4\pi d}{\lambda}\right) \tag{19.71}$$

where t is the time, T is the averaging interval, and A is the amplitude factor from gain of integrating amplifier.

From the earlier expression for ϕ, it can be seen that the quantity actually measured is in fact the *cosine* of the phase shift and not the phase shift itself (Woodbury et al., 1993). This situation introduces a so-called *ambiguity interval* for scenarios where the round-trip distance exceeds the modulation wavelength λ (i.e., the phase measurement becomes ambiguous once ϕ exceeds 360°). Conrad and Sampson (1990) define this ambiguity interval as the maximum range that allows the phase difference to go through one complete cycle of 360°:

$$R_a = \frac{c}{2f} \tag{19.72}$$

where R_a is the ambiguity range interval.

Referring to Eq. (19.73), it can be seen that the total round-trip distance $2d$ is equal to some integer number of wavelengths $n\lambda$ plus the fractional wavelength distance x associated with the phase shift. Since the cosine relationship is not single-valued for all of ϕ, there will be more than one distance d corresponding to any given phase-shift measurement (Woodbury et al., 1993):

$$\cos\phi = \cos\left(\frac{4\pi d}{\lambda}\right) = \cos\left(\frac{2\pi(x + n\lambda)}{\lambda}\right) \tag{19.73}$$

where

$d = (x + n\lambda)/2$ = true distance to target,
x = distance corresponding to differential phase ϕ,
n = number of complete modulation cycles.

Careful re-examination of Eq. (19.73), in fact, shows that the cosine function is not single-valued even within a solitary wavelength interval of 360°. Accordingly, if only the cosine of the phase angle is measured, the ambiguity interval must be further reduced to half the modulation wavelength, or 180° (Scott, 1990). In addition, the slope of the curve is such that the rate of change of the nonlinear cosine function is not constant over the range of $0 \leq \phi \leq 180°$, and is in fact zero at either extreme. The achievable accuracy of the phase-shift measurement technique thus varies as a function of target distance, from best-case

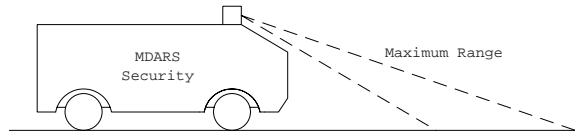



FIGURE 19.70 By limiting the maximum distance measured to be less than the range *ambiguity interval* R_a, erroneous distance measurements can be avoided.

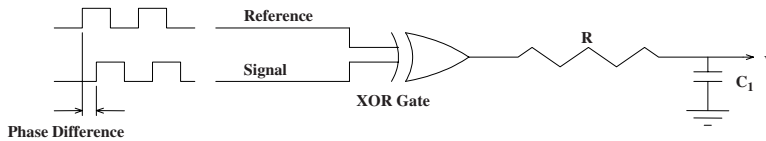



FIGURE 19.71 At low frequencies typical of ultrasonic systems, a simple phase-detection circuit based on an *exclusive-or* gate will generate an analog output voltage proportional to the phase difference seen by the inputs (adapted from Figueroa & Barbieri, 1991a).

performance for a phase angle of 90° to worst case at 0 and 180°. For this reason, the useable measurement range is typically even further limited to 90% of the 180° ambiguity interval (Chen et al., 1993).

A common solution to this problem involves taking a second measurement of the same scene but with a 90° phase shift introduced into the reference waveform, the net effect being the sine of the phase angle is then measured instead of the cosine. This additional information (i.e., both sine and cosine measurements) can be used to expand the phase angle ambiguity interval to the full 360° limit previously discussed (Scott, 1990). Furthermore, an overall improvement in system accuracy is achieved, as for every region where the cosine measurement is insensitive (i.e., zero slope), the complementary sine measurement will be at peak sensitivity (Woodbury et al., 1993).

Nevertheless, the unavoidable potential for erroneous information as a result of the ambiguity interval is a detracting factor in the case of phase-detection schemes. Some applications simply avoid such problems by arranging the optical path in such a fashion as to ensure the maximum possible range is always less than the ambiguity interval (Fig. 19.70). Alternatively, successive measurements of the same target using two different modulation frequencies can be performed, resulting in two equations with two unknowns, allowing both x and n (in the previous equation) to be uniquely determined. Kerr (1988) describes such an implementation using modulation frequencies of 6 and 32 MHz.

For square-wave modulation at the relatively low frequencies typical of ultrasonic systems (20–200 kHz), the phase difference between incoming and outgoing waveforms can be measured with the simple linear circuit shown in Fig. 19.71 (Figueroa & Barbieri, 1991a). The output of the *exclusive-or* gate goes high whenever its inputs are at opposite logic levels, generating a voltage across capacitor C_1 that is proportional to the phase shift. For example, when the two signals are in phase (i.e., $\phi = 0$), the gate output stays low and V is zero; maximum output voltage occurs when ϕ reaches 180°. While easy to implement, this simplistic approach is limited to very low frequencies and may require frequent calibration to compensate for drifts and offsets due to component aging or changes in ambient conditions (Figueroa & Lamancusa, 1992).

Extended Range Phase Measurement Systems

Figueroa and Barbieri (1991a; 1991b) report an interesting method for extending the ambiguity interval in ultrasonic phase-detection systems through frequency division of the received and reference signals. Since the span of meaningful comparison is limited (best case) to one wavelength, λ, it stands to reason that decreasing the frequency of the phase detector inputs by some common factor will increase λ by a similar amount. The concept is illustrated in Fig. 19.72. Due to the very short wavelength of ultrasonic energy (i.e., about 0.25 in. for the Polaroid system at 49.1 kHz), the total effective range is still only 4 in.

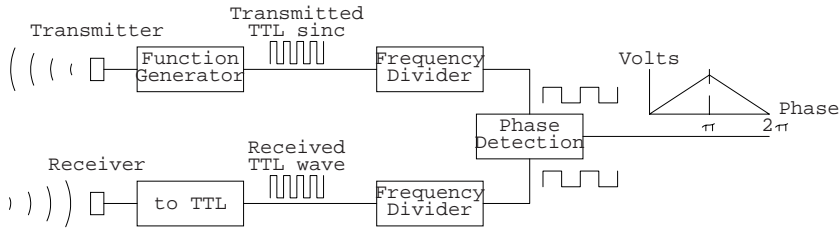



FIGURE 19.72 Dividing the input frequencies to the phase comparator by some common integer value will extend the ambiguity interval by the same factor, at the expense of resolution (adapted from Figueroa & Barbieri, 1991a).

after dividing the detector inputs by a factor of 16. Due to this inherent range limitation, ultrasonic phase-detection ranging systems are not extensively applied in mobile robotic applications, although Figueroa and Lamancusa (1992) describe a hybrid approach used to improve the accuracy of TOF ranging for three-dimensional position location.

An ingenious method to measure range using phase information was developed by Young and Li (1992). The method reconstructs the total range by piecing together multiple consecutive phase chunks that reset every 2π radians of phase difference between emitted and received signals. This is another method that overcomes the limitation of phase-based systems to ranges shorter than one acoustic wavelength. The discontinuities at every 2π radians are eliminated by first taking the derivative of the phase, resulting in a smooth signal with sharp pulses (impulses) at the location of each discontinuity. Subsequently, the pulses are ignored and the result is integrated and multiplied by a constant to reconstruct the overall range. The method was tested with an experiment that employed 40 kHz transducers. Distances from 40 to 400 mm were measured with errors from ±0.1629 to ±0.4283 mm.

Laser-based continuous-wave ranging originated out of work performed at the Stanford Research Institute in the 1970s (Nitzan et al., 1977). Range accuracies approach those achievable by pulsed laser TOF methods. Only a slight advantage is gained over pulsed TOF rangefinding, however, since the difficult time-measurement problem is replaced by the need for fairly sophisticated phase-measurement electronics (Depkovich & Wolfe, 1984). In addition, problems with the phase-shift measurement approach are routinely encountered in situations where the outgoing energy is simultaneously reflected from two target surfaces at different distances from the sensor, as for example when scanning past a prominent vertical edge (Hebert & Krotkov, 1991).

The system electronics are set up to compare the phase of a single incoming wave with that of the reference signal and are not able to cope with two superimposed reflected waveforms. Adams (1993) describes a technique for recognizing the occurrence of this situation in order to discount the resulting erroneous data.

Frequency Modulation

This is a method devised to improve the accuracy in detecting the time-of-arrival of the wave to the receiver. Instead of a single frequency wave, a frequency modulated wave of the form $f = f_0 + kt$ is emitted. The difference between the emitted and received frequency at any time is $\Delta f = kt - k(t - t_f) = kt_f$ (Fig. 19.73). The advantage of this method is that one does not need to know exactly when the wave arrived to the receiver. However, accurate real-time frequency measurement electronics must be used, and the transducers must respond within the frequency band sweep. Modulation other than linear is also possible in order to improve signal-to-noise ratio and hence accuracy.

The signal is reflected from a target and arrives at the receiver at time $t + T$:

$$T = \frac{2d}{c}$$

where T is the round-trip propagation time, d is the distance to target, and c is the speed of travel.

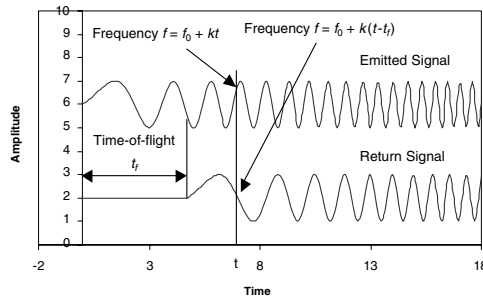



FIGURE 19.73 The frequency difference between the emitted signal and received signal is proportional to the time-of-flight at any given time.

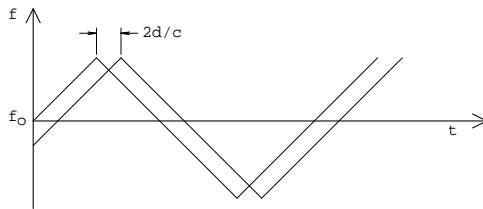



FIGURE 19.74 The received frequency curve is shifted along the time axis relative to the reference frequency.

The received signal is compared with a reference signal taken directly from the transmitter. The received frequency curve (Fig. 19.74) will be displaced along the time axis relative to the reference frequency curve by an amount equal to the time required for wave propagation to the target and back. (There might also be a vertical displacement of the received waveform along the frequency axis, due to the Doppler effect.) These two frequencies when combined in the mixer produce a beat frequency F_b:

$$F_b = f(t) - f(T + t) = kT$$

where k is a constant.

This beat frequency is measured and used to calculate the distance to the object:

$$d = \frac{F_b c}{4F_r F_d}$$

where

d = range to target,
c = speed of light,
F_b = beat frequency,
F_r = repetition (modulation) frequency,
F_d = total FM frequency deviation.

Distance measurement is therefore directly proportional to the difference or beat frequency and is as accurate as the linearity of the frequency variation over the counting interval.

Advances in wavelength control of laser diodes now permit this ultrasonic and radar ranging technique to be used with lasers. The frequency or wavelength of a laser diode can be shifted by varying its temperature. Consider an example where the wavelength of an 850-nm laser diode is shifted by 0.05 nm in 4 μs: the corresponding frequency shift is 5.17 MHz/ns. This laser beam, when reflected from a surface

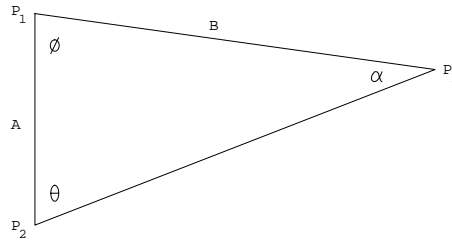



FIGURE 19.75 Triangulation ranging systems determine range B to target point P_3 by measuring angles *f* and *q* at points P_1 and P_2.

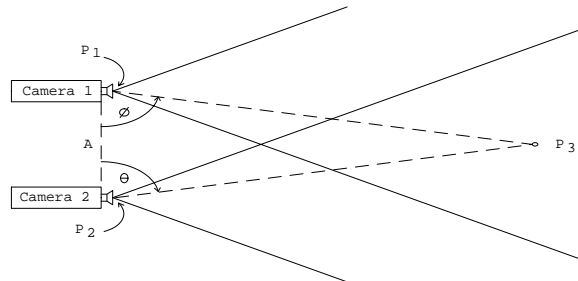



FIGURE 19.76 Passive stereoscopic ranging system configuration.

1 m away, would produce a beat frequency of 34.5 MHz. The linearity of the frequency shift controls the accuracy of the system.

The frequency-modulation approach has an advantage over the phase-shift measurement technique in which a single distance measurement is not ambiguous. (Recall that phase-shift systems must perform two or more measurements at different modulation frequencies to be unambiguous.) However, frequency modulation has several disadvantages associated with the required linearity and repeatability of the frequency ramp, as well as the coherence of the laser beam in optical systems. As a consequence, most commercially available FMCW ranging systems are radar based, while laser devices tend to favor TOF and phase-detection methods.

Triangulation Ranging

Triangulation ranging is based upon an important premise of plane trigonometry, which states that given the length of a side and two angles of a triangle, it is possible to determine the length of the other sides and the remaining angle. The basic *Law of Sines* can be rearranged as shown below to represent the length of side B as a function of side A and the angles θ and ϕ:

In ranging applications, length B would be the desired distance to the object of interest at point P_3 (Fig. 19.75) for known sensor separation baseline A.

Triangulation ranging systems are classified as either *passive* (use only the ambient light of the scene) or *active* (use an energy source to illuminate the target). Passive stereoscopic ranging systems position directional detectors (video cameras, solid-state imaging arrays, or position sensitive detectors) at positions corresponding to locations P_1 and P_2 (Fig. 19.76). Both imaging sensors are arranged to view the same object point, P_3, forming an imaginary triangle. The measurement of angles θ and ϕ in conjunction with the known orientation and lateral separation of the cameras allows the calculation of range to the object of interest.

Active triangulation systems, on the other hand, position a controlled light source (such as a laser) at either point P_1 or P_2, directed at the observed point P_3. A directional imaging sensor is placed at the remaining triangle vertex and is also aimed at P_3. Illumination from the source will be reflected by the

target, with a portion of the returned energy falling on the detector. The lateral position of the spot as seen by the detector provides a quantitative measure of the unknown angle ϕ, permitting range determination by the *Law of Sines*.

The performance characteristics of triangulation systems are to some extent dependent on whether the system is active or passive. Passive triangulation systems using conventional video cameras require special ambient lighting conditions that must be artificially provided if the environment is too dark. Furthermore, these systems suffer from a correspondence problem resulting from the difficulty in matching points viewed by one image sensor with those viewed by the other. On the other hand, active triangulation techniques employing only a single detector do not require special ambient lighting, nor do they suffer from the correspondence problem. Active systems, however, can encounter instances of no recorded strike because of specular reflectance or surface absorption of the light.

Limiting factors common to all triangulation sensors include reduced accuracy with increasing range, angular measurement errors, and a *missing parts* (also known as *shadowing*) problem. *Missing parts* refers to the scenario where particular portions of a scene can be observed by only one viewing location (P_1 or P_2). This situation arises because of the offset distance between P_1 and P_2, causing partial occlusion of the target (i.e., a point of interest is seen in one view but otherwise occluded or not present in the other). The design of triangulation systems must include a tradeoff analysis of the offset: as this baseline measurement increases, the range accuracy increases, but problems due to directional occlusion worsen.

Stereo Disparity

The first of the triangulation schemes to be discussed, *stereo disparity* (also called *stereo vision, binocular vision*, and *stereopsis*) is a passive ranging technique modeled after the biological counterpart. When a three-dimensional object is viewed from two locations on a plane normal to the direction of vision, the image as observed from one position is shifted laterally when viewed from the other. This displacement of the image, known as *disparity*, is inversely proportional to the distance to the object. Humans subconsciously *verge* their eyes to bring objects of interest into rough registration (Burt et al., 1992). Hold up a finger a few inches away from your face while focusing on a distant object and you can simultaneously observe two displaced images in the near field. In refocusing on the finger, your eyes actually turn inward slightly to where their respective optical axes converge at the finger instead of infinity.

Most implementations use a pair of identical video cameras (or a single camera with the ability to move laterally) to generate the two disparity images required for stereoscopic ranging. The cameras are typically aimed straight ahead viewing approximately the same scene, but (in simplistic cases anyway) do not possess the capability to *verge* their center of vision on an observed point, as can human eyes. This limitation makes placement of the cameras somewhat critical because stereo ranging can take place only in the region where the fields of view overlap. In practice, analysis is performed over a selected range of disparities along the Z axis on either side of a perpendicular plane of zero disparity called the *horopter* (Fig. 19.77). The selected image region in conjunction with this disparity range defines a three-dimensional volume known as the *stereo observation window* (Burt et al., 1993).

More recently there has evolved a strong interest within the research community for dynamically reconfigurable camera orientation (Fig. 19.78), often termed *active vision* in the literature (Aloimonos et al., 1987; Swain & Stricker, 1991; Wavering et al., 1993). The widespread acceptance of this terminology is perhaps somewhat unfortunate in view of potential confusion with stereoscopic systems employing an active illumination source (see section 4.1.3). *Verging stereo*, another term in use, is perhaps a more appropriate choice. *Mechanical verging* is defined as the process of rotating one or both cameras about the vertical axis in order to achieve zero disparity at some selected point in the scene (Burt et al., 1992).

There are four basic steps involved in the stereo ranging process (Poggio, 1984):

- A point in the image of one camera must be identified (Fig. 19.79, left).
- The same point must be located in the image of the other camera (Fig. 19.79, right).
- The lateral positions of both points must be measured with respect to a common reference.
- Range Z is then calculated from the disparity in the lateral measurements.

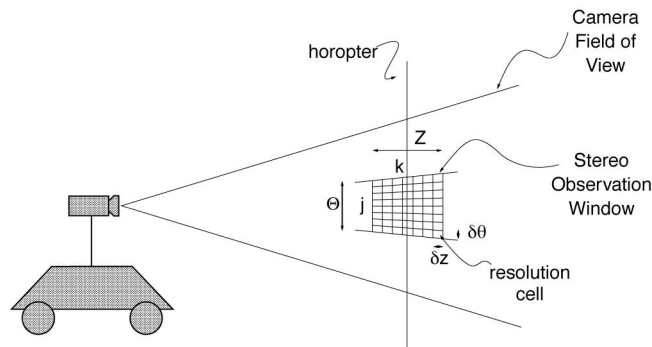



FIGURE 19.77 The *stereo observation window* is that volume of interest on either side of the plane of zero disparity known as the *horopter* (courtesy David Sarnoff Research Center).

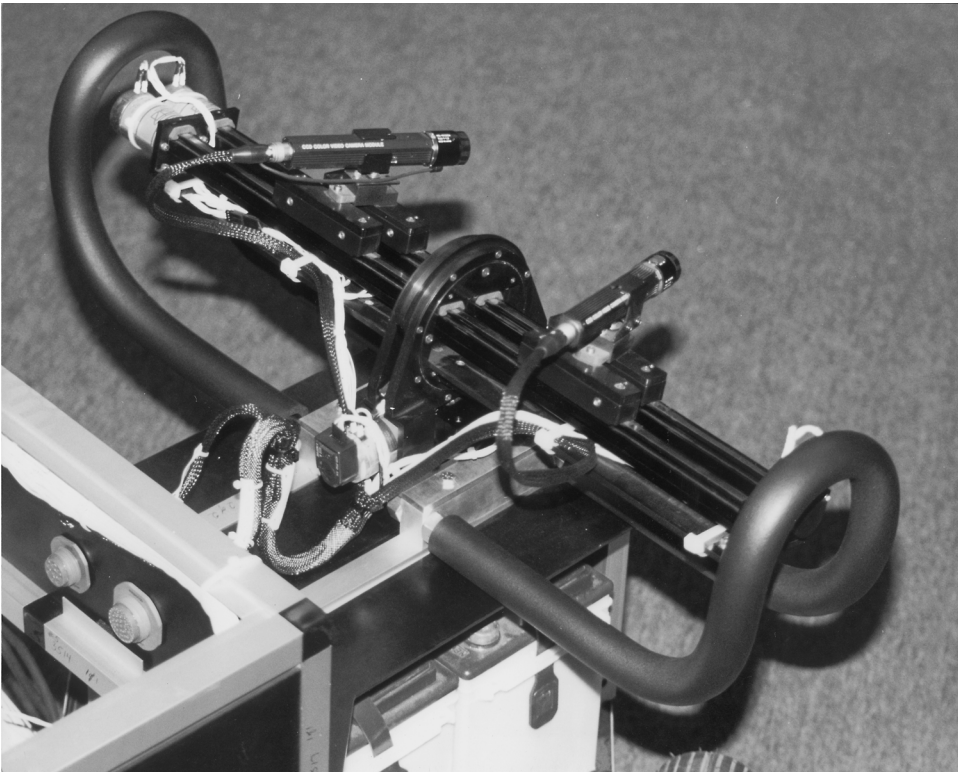



FIGURE 19.78 This stereoscopic camera mount uses a pair of lead-screw actuators to provide reconfigurable baseline separation and vergence as required (courtesy Robotic Systems Technology, Inc.).

On the surface this procedure appears rather straightforward, but difficulties arise in practice when attempting to locate the specified point in the second image (Fig. 19.79). The usual approach is to match "interest points" characterized by large intensity discontinuities (Conrad & Sampson, 1990). Matching is complicated in regions where the intensity and/or color are uniform (Jarvis, 1983b). Additional factors include the presence of shadows in only one image (due to occlusion) and the variation in image characteristics that can arise from viewing environmental lighting effects from different angles. The effort to match the two images of the point is called *correspondence*, and methods for minimizing this

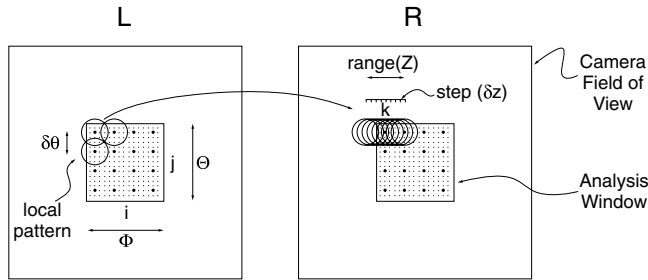



FIGURE 19.79 Range Z is derived from the measured disparity between interest points in the left and right camera images (courtesy David Sarnoff Research Center).

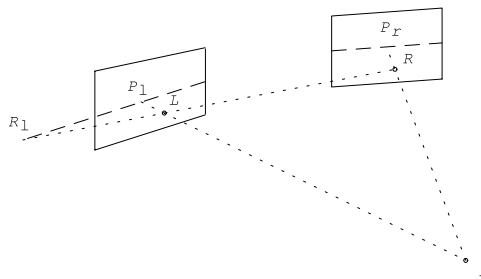



FIGURE 19.80 The *epipolar surface* is a plane defined by the lens centerpoints L and R and the object of interest at P (adapted from Vuylsteke et al., 1990).

computationally expensive procedure are widely discussed in the literature (Nitzan, 1981; Jarvis, 1983a; Poggio, 1984; Loewenstein, 1984; Vuylsteke et al., 1990; Wildes, 1991).

Probably the most basic simplification employed in addressing the otherwise overwhelming *correspondence* problem is seen in the *epipolar restriction* that reduces the two-dimensional search domain to a single dimension (Vuylsteke et al., 1990). The *epipolar surface* is a plane defined by the point of interest P and the positions of the left and right camera lenses at L and R, as shown in Fig. 19.80. The intersection of this plane with the left image plane defines the *left epipolar line* as shown. As can be seen from the diagram, since the point of interest P lies in the *epipolar plane*, its imaged point P_l must lie somewhere along the *left epipolar line*. The same logic dictates that the imaged point P_r must lie along a similar *right epipolar line* within the right image plane. By carefully aligning the camera image planes such that the *epipolar lines* coincide with identical scan lines in their respective video images, the correspondence search in the second image is constrained to the same horizontal scan line containing the point of interest in the first image. This effect can also be achieved with nonaligned cameras by careful calibration and rectification (resampling).

To reduce the image processing burden, most correspondence schemes monitor the overall scene at relatively low resolution and examine only selected areas in greater detail. A *foveal representation* analogous to the acuity distribution in human vision is generally employed as illustrated in Fig. 19.81, allowing an extended field-of-view without loss of resolution or increased computational costs (Burt et al., 1993). The high-resolution *fovea* must be shifted from frame to frame in order to examine different regions of interest individually. Depth acuity is greatest for small disparities near the horopter and falls off rapidly with increasing disparities (Burt et al., 1992).

Active Triangulation

Rangefinding by *active triangulation* is a variation on the *stereo disparity* method of distance measurement. In place of one camera is a laser (or LED) light source aimed at the surface of the object of interest.

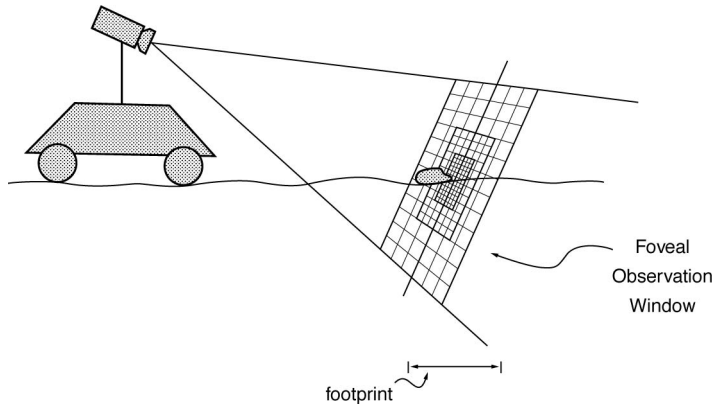



FIGURE 19.81 The *foveal* stereo representation provides high acuity near the center of the *observation window*, with decreasing resolution towards the periphery (Courtesy David Sarnoff Research Center).

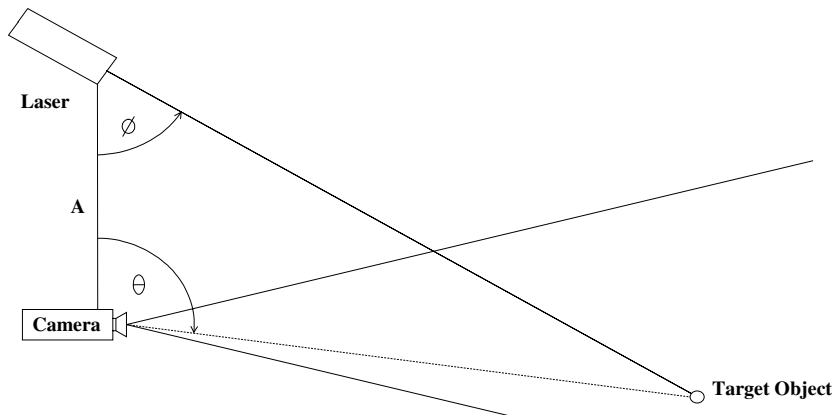



FIGURE 19.82 An *active triangulation-ranging configuration* employing a conventional CCD array as the detector.

The remaining camera is offset from this source by a known distance *A* and configured to hold the illuminated spot within its field of view (Fig. 19.82).

For one- or two-dimensional array detectors such as vidicon or CCD cameras, the range can be determined from the known baseline distance *A* and the relative position of the laser-spot image on the image plane. For mechanically scanned single-element detectors such as photodiodes or phototransistors, the rotational angles of the detector and/or source are measured at the exact instant the detector observes the illuminated spot. The trigonometric relationships between these angles and the baseline separation are used (in theory) to compute the distance. To obtain three-dimensional information for a volumetric region of interest, laser triangulators can be scanned in both azimuth and elevation. In systems where the source and detector are self-contained components, the entire configuration can be moved mechanically. In systems with movable optics, the mirrors and lenses are generally scanned in synchronization while the laser and detector remain stationary.

Drawbacks to active triangulation include the *missing parts* situation, where points illuminated by the light source cannot be seen by the camera and vice versa (Jarvis, 1983b), as well as surface absorption or specular reflection of the irradiating energy (see Chapter 9). On the positive side, however, point-source illumination of the image effectively eliminates the correspondence problem encountered in stereo disparity rangefinders. There is also no dependence on scene contrast, and reduced influence from ambient lighting effects. (Background lighting is effectively a noise source that can limit range resolution.)

FIGURE 19.83 HERMIES IIB employed an active stereoscopic ranging system with an external laser source that could be used to designate objects of interest in the video image (courtesy Oak Ridge National Laboratory).

Active Stereoscopic

Due to the computationally intensive complexities and associated resources required for establishing correspondence, passive stereoscopic methods were initially limited in practical embodiments to very simple scenes (Blais et al., 1988). One way around these problems is to employ an active source in conjunction with a pair of stereo cameras. This active illumination greatly improves system performance when viewing scenes with limited contrast. Identification of the light spot becomes a trivial matter; a video frame representing a scene illuminated by the source is subtracted from a subsequent frame of the same image with the light source deactivated. Simple thresholding of the resultant difference image quickly isolates the region of active illumination. This process is performed in rapid sequence for both cameras, and the lateral displacement of the centroid of the spot is then determined.

Alignment between the source and cameras is not critical in active stereoscopic ranging systems; in fact, the source does not even have to be located on board the robot. For example, Kilough and Hamel (1989) describe two innovative configurations using external sources for use with the robot HERMIES IIB, built at Oak Ridge National Laboratory. A pair of wide-angle black-and-white CCD cameras are mounted on a pan-and-tilt mechanism atop the robot's head, as shown in Fig. 19.83. Analog video outputs from the cameras are digitized by a frame grabber into a pair of 512 by 384-pixel arrays, with offboard image processing performed by a *Hypercube* at a scaled-down resolution of 256 by 256. The initial application of the vision system was to provide control of a pair of robotic arms (from the Heathkit *HERO-1* robot) employed on HERMIES.

To accomplish this task, a near-infrared LED is attached to the end of the HERO-1 arm near the manipulator and oriented so as to be visible within the field of view of the stereo camera pair. A sequence of images is then taken by each camera, with the LED first *on* and then *off*. The *off* representations are subtracted from the *on* representations, leaving a pair of difference images, each comprised of a single bright dot representing the location of the LED. The centroids of the dots are calculated to

precisely determine their respective coordinates in the difference-image arrays. A range vector to the LED can then be easily calculated, based on the lateral separation of the dots as perceived by the two cameras. This technique establishes the actual location of the manipulator in the reference frame of the robot. Experimental results indicated a 2-in. accuracy with a 0.2-in. repeatability at a distance of approximately 2 ft (Kilough and Hamel, 1989).

A near-infrared solid-state laser mounted on a remote tripod was then used by the operator to designate a target of interest within the video image of one of the cameras. The same technique described above was repeated, only this time the imaging system toggled the laser power *on* and *off*. A subsequent differencing operation enabled calculation of a range vector to the target, also in the robot's reference frame. The difference in location of the gripper and the target object could then be used to effect both platform and arm motion. The imaging processes would alternate in near-real-time for the gripper and the target, enabling the HERMIES robot to drive over and grasp a randomly designated object under continuous closed-loop control.

Structured Light

Ranging systems that employ *structured light* are a further refined case of active triangulation. A pattern of light (either a line, a series of spots, or a grid pattern) is projected onto the object surface while the camera observes the pattern from its offset vantage point. Range information manifests itself in the distortions visible in the projected pattern due to variations in the depth of the scene. The use of these special lighting effects tends to reduce the computational complexity and improve the reliability of three-dimensional object analysis (Jarvis, 1983b; Vuylsteke et al., 1990). The technique is commonly used for rapid extraction of limited quantities of visual information of moving objects (Kent, 1985), and thus lends itself well to collision avoidance applications. Besl (1988) provides a good overview of *structured-light* illumination techniques, while Vuylsteke et al. (1990) classify the various reported implementations according to the following characteristics:

- The number and type of sensors
- The type of optics (i.e., spherical or cylindrical lens, mirrors, multiple apertures)
- The dimensionality of the illumination (i.e., point or line)
- Degrees of freedom associated with scanning mechanism (i.e., zero, one, or two)
- Whether or not the scan position is specified (i.e., the instantaneous scanning parameters are not needed if a redundant sensor arrangement is incorporated)

The most common *structured-light* configuration entails projecting a line of light onto a scene, originally introduced by P. Will and K. Pennington of IBM Research Division Headquarters, Yorktown Heights, NY (Schwartz, undated). Their system created a plane of light by passing a collimated incandescent source through a slit, thus projecting a line across the scene of interest. (More recent systems create the same effect by passing a laser beam through a cylindrical lens or by rapidly scanning the beam in one dimension.) Where the line intersects an object, the camera view will show displacements in the light stripe that are proportional to the depth of the scene. In the example depicted in Fig. 19.84, the lower the reflected illumination appears in the video image, the closer the target object is to the laser source. The exact relationship between stripe displacement and range is dependent on the length of the baseline

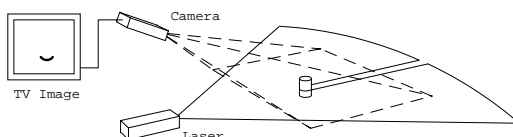



FIGURE 19.84 A common structured-light configuration used on robotic vehicles projects a horizontal line of illumination onto the scene of interest and detects any target reflections in the image of a downward-looking CCD array.

between the source and the detector. Like any triangulation system, when the baseline separation increases, the accuracy of the sensor increases, but the *missing parts* problem worsens.

Three-dimensional range information for an entire scene can be obtained in relatively simple fashion through striped lighting techniques. By assembling a series of closely spaced two-dimensional contours, a three-dimensional description of a region within the camera's field of view can be constructed. The third dimension is typically provided by scanning the laser plane across the scene. Compared to single-point triangulation, striped lighting generally requires less time to digitize a surface, with fewer moving parts because of the need to mechanically scan only in one direction. The drawback to this concept is that range extraction is time consuming and difficult due to the necessity of storing and analyzing many frames.

An alternative structured-light approach for three-dimensional applications involves projecting a rectangular grid of high-contrast light points or lines onto a surface. Variations in depth cause the grid pattern to distort, providing a means for range extraction. The extent of the distortion is ascertained by comparing the displaced grid with the original projected patterns as follows (LeMoigue & Waxman, 1984):

- Identify the intersection points of the distorted grid image.
- Label these intersections according to the coordinate system established for the projected pattern.
- Compute the disparities between the intersection points and/or lines of the two grids.
- Convert the displacements to range information.

The comparison process requires correspondence between points on the image and the original pattern, which can be troublesome. By correlating the image grid points to the projected grid points, this problem can be somewhat alleviated. A critical design parameter is the thickness of the lines that make up the grid and the spacing between these lines. Excessively thin lines will break up in busy scenes, causing discontinuities that adversely affect the intersection points labeling process. Thicker lines will produce less observed grid distortion resulting in reduced range accuracy (LeMoigue & Waxman, 1984). The sensor's intended domain of operation will determine the density of points required for adequate scene interpretation and resolution.

Magnetic Position Measurement Systems

Magnetic tracking uses a source element radiating a magnetic field (three axes) and a small sensor (three axes) that reports its position and orientation with respect to the source. Competing systems provide various multi-source, multi-sensor systems that will track a number of points at up to 100 Hz in ranges from 3 to 20 ft (Polhemus Incorporated, and Ascension Technologies). They are generally accurate to better than 0.1 in. in position and 0.1° in rotation. Magnetic systems do not rely on line-of-sight from source to object, as do optical and acoustic systems, but metallic objects in the environment will distort the magnetic field, giving erroneous readings. They require cable attachment to a central device (as do LEDs and acoustic systems). Current technology is quite robust and widely used for single or double hand-tracking, head-mounted devices, biomechanical analysis, graphics (digitization in 3D), stereotaxic localization, etc.

Magnetic field sources can be AC or DC. DC sources may emit pulses rather than continuous radiation in order to minimize interference from other magnetic sources. Using pulsed systems allows measurement of existing magnetic fields in the environment during the inactive period. Knowledge of these magnetic fields external to the system is used to improve accuracy and to overcome sensitivity to metals.

Figure 19.85 shows a typical transmitter-drive electronics (courtesy of Ascension Technologies). It provides DC current pulses to each antenna of the transmitter, one antenna at a time. The transmitter consists of a core about which the X, Y, and Z antennae are wound. While a given transmitter antenna is activated with current, readings are taken from all three antennae of the sensor. Initially the transmitter is shut off so that the sensor can measure the x, y, and z components of the earth's magnetic field. During operation, the computer sends to the digital-to-analog (D/A) converter a number that represents the amplitude of the current pulse to be sent to the selected transmitter antenna. The D/A converter converts

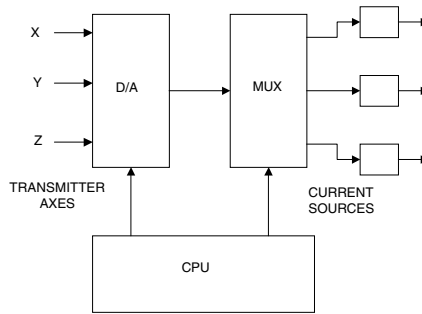



FIGURE 19.85 Magnetic Positioning System: Transmitting Circuit (Ascension Technologies).

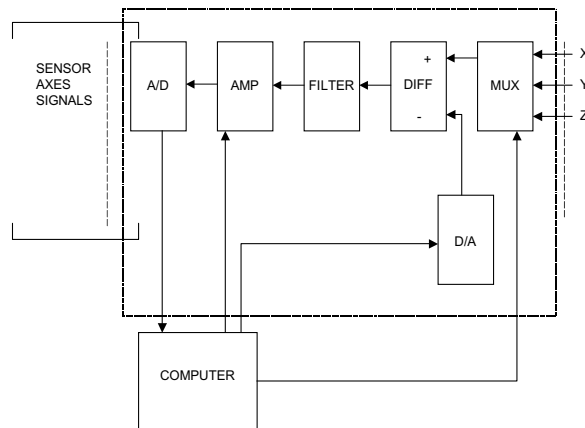



FIGURE 19.86 Magnetic Positioning System: Receiving Circuit (Ascension Technologies).

this amplitude to an analog control voltage. This control voltage goes to the multiplexer (MUX), which connects it to the X, Y, or Z transmitter current source.

The sensor consists of three orthogonal antennae sensitive to DC magnetic fields. Many technologies can be used to implement the DC sensor. The Flock (Ascension Technologies) uses a three-axis fluxgate magnetometer. The output from the sensor goes to the signal processing electronics. As detailed in Fig. 19.86, the sensor signal processing electronics consists of a multiplexer (MUX), which, on command from the computer, switches the desired X, Y, or Z sensor antenna signal, one at a time, to the differential amplifier (DIFF). The differential amplifier subtracts from this antenna signal the previously measured component of the earth's magnetic field. It outputs only that part of the received signal that is due to the transmitted field. The output from the differential amplifier is then filtered to remove noise and amplified. The analog-to-digital converter converts the DC signal to a digital format that can be read by the computer.

Other Distance Measuring Methods

The following methods are used to measure displacement, and thus can be used to infer distance travelled for certain applications.

Odometry

This is one of many methods to measure position and it is an indirect method of determining range. Range is determined by measuring the rotation of a wheel as it traverses from the reference to the target location. Wheel rotation is measured using angular encoders that may be digital or analog in nature.

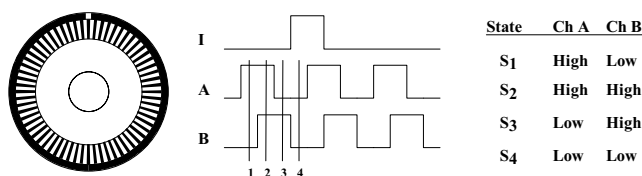



FIGURE 19.87 Optical Incremental Angular Encoder.

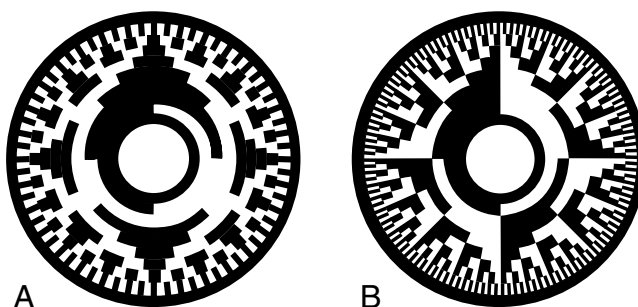



FIGURE 19.88 Absolute Optical Encoder.

Angular Optical Encoders

These devices encompass a light source, optics to shape and guide the light, a coded wheel with transparent and opaque sections, and a light detector array. There are two types of optical encoders: incremental and absolute.

Incremental Angular Optical Encoders. A schematic is shown in Fig. 19.87. The wheel is opaque except for the slots along the circumference. Two rows of slots displaced by a 90° phase are used to determine rotation and direction. As the wheel rotates, two pulse chains 90° out of phase with each other are generated (Channels A and B). Distance is determined by counting the number of pulses (and quarter pulses for increased resolution) from the initial arbitrary zero position. The reference is lost when power is interrupted. The direction of motion is defined by determining which pulse chain leads the other.

An index pulse that appears once per revolution is also usually available (Channel Z). The resolution depends upon the number of slots around the circumference. Larger wheels can accommodate higher resolution. The total distance traveled depends upon the system used to count pulses. Since pulse counting is done outside the sensor, distances in the meters may be measured.

Decoding of the pulses (Channels A, B, and Z) to obtain angular displacement is done using specialized chips (Hewlett Packard makes a family of chips), or full-fledged integrated circuit boards. Data acquisition boards with counters may also be programmed to decode range from these sensors.

Commercially available units have a maximum operating speed of about 6000 rpm, maximum counts per revolution of about 360,000, a maximum resolution of 0.001°, and a frequency response of up to 150 kHz.

Absolute Angular Optical Encoders. These encoders have a wheel which is coded in such a way that each angular slot represents a number of bits that may be either *on* (transparent) or *off* (opaque) (Fig. 19.88). Therefore, the angular position of the wheel has an absolute value given by the code of the angular slot currently aligned with the optics. Its position is known even after turning the power *off* and *on* again. When the optics crosses the line between two slots, the pattern changes to indicate an increment of one unit. However, the position is uncertain if the wheel stops with the optics right on the line. To decrease this uncertainty, the patterns are defined according to the Grey Code. In this coding scheme, only one bit of the pattern changes from one slot to the next. Thus, the uncertainty is minimized to one unit.

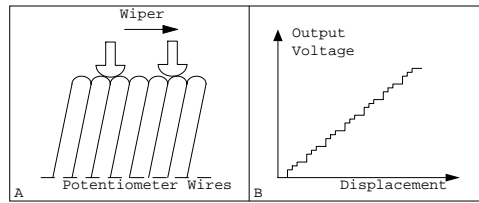



FIGURE 19.89 Potentiometer: Principle of Operation.

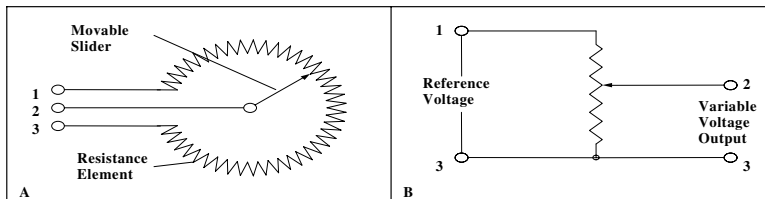



FIGURE 19.90 Potentiometer: Circuit Representation.

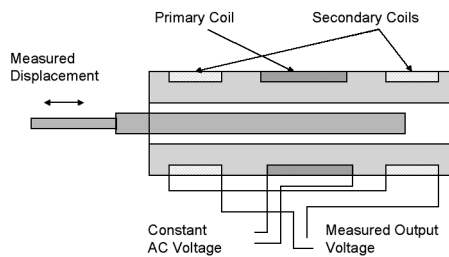



FIGURE 19.91 Linear Variable Differential Transformer.

These sensors are suitable to measure small ranges, in the order of hundreds of millimeters. Larger wheels can accommodate more slots and more bits per slot. Commercial units of 11 bits are available, with a resolution of ±1/2 of the least significant bit, and a frequency response of 100 K 11-bit words per second.

Linear Optical Encoders. These are the same as angular encoders, except that instead of a coded wheel, they have a coded bar and a slider that carries the optical and electronic components. Distance is measured along the bar. In commercial units, the maximum measuring distance is about 2.150 m, the maximum resolution 0.08 μm, and the maximum operating speed 508 mm/s.

Potentiometers

Potentiometers are variable electrical resistance transducers. They consist of a winding and a sliding contact. As the sliding contact moves along the winding, the resistance changes in linear relationship with the distance from one end of the potentiometer (Fig. 19.89). The variable resistance is wired as a voltage divider so that the output voltage is proportional to the distance traveled by the wiper (Fig. 19.90). The resolution is defined by the number of turns per unit distance, and loading effects of the voltage divider circuit should be considered.

Linear Variable Differential Transformers

The linear variable differential transformer (LVDT) generates an AC signal whose magnitude is related to the displacement of a moving core (Fig. 19.91). As the core changes position with respect to the coils, it changes the magnetic field, and thence the voltage amplitude in the secondary coil.

LVDT resolution depends on the instruments used to measure voltage. 25-μm resolution can be achieved. Stationary (low frequency) signals may be measured using an AC meter. High frequency signals require specialized electronics for demodulation or a data acquisition system to process the signal using a PC.

A rotary variable differential transformer (RVDT) operates under the same principle as the LVDT and is available with a range of approximately ±40°.

Proximity Sensors

Proximity sensors, used to determine the presence (as opposed to actual range) of nearby objects, were developed to extend the sensing range beyond that afforded by direct-contact tactile or haptic sensors. Recent advances in electronic technology have significantly improved performance and reliability, thereby increasing the number of possible applications. As a result, many industrial installations that historically have used mechanical limit switches can now choose from a variety of alternative noncontact devices for their close (between a fraction of an inch and a few inches) sensing needs. Such *proximity sensors* are classified into several types in accordance with the specific properties used to initiate a switching action:

- Magnetic
- Inductive
- Ultrasonic
- Microwave
- Optical
- Capacitive

The reliability characteristics displayed by these sensors make them well suited for operation in harsh or otherwise adverse environments, while providing high-speed response and long service lives. Instruments can be designed to withstand significant shock and vibration, with some capable of handling forces over 30,000 Gs and pressures of nearly 20,000 psi (Hall, 1984). Burreson (1989) and Peale (1992) discuss advantages and tradeoffs associated with proximity sensor selection for applications in challenging and severe environments. In addition, proximity devices are valuable when detecting objects moving at high speed, when physical contact may cause damage, or when differentiation between metallic and nonmetallic items is required. Ball (1986), Johnson (1987), and Wojcik (1994) provide general overviews of various alternative proximity sensor types with suggested guidelines for selection.

Magnetic Proximity Sensors

The simplest form of magnetic proximity sensor is the *magnetic reed switch*, schematically illustrated in Fig. 19.92. A pair of low-reluctance ferromagnetic reeds are cantilevered from opposite ends of a hermetically sealed tube, arranged such that their tips overlap slightly without touching. The extreme ends of the reeds assume opposite magnetic polarities when exposed to an external magnetic flux, and the subsequent attractive force across the gap pulls the flexible reed elements together to make electrical contact (Hamlin, 1988).

Available in both *normally open* and *normally closed* configurations, these inexpensive and robust devices are commonly employed as door- and window-closure sensors in security applications. Some problems

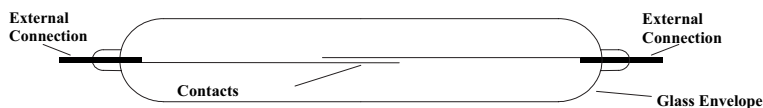



FIGURE 19.92 The hermetically sealed *magnetic reed switch*, shown here with normally open contacts, is filled with inert gas and impervious to dust and corrosion.

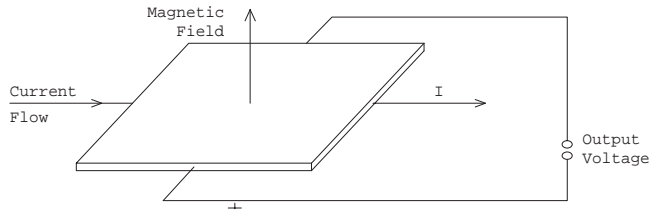



FIGURE 19.93 In 1879, E.H. Hall discovered a small transverse voltage was generated across a current-carrying conductor in the presence of a static magnetic field, a phenomenon now known as the *Hall effect* (adapted from Lenz, 1990).

can be encountered with this type of sensor due to contact bounce, structural vibration, and pitting of the mating surfaces in the case of inductive or capacitive loads (Burreson, 1989), prompting most designers to opt instead for the more reliable solid-state Hall-effect magnetic sensor.

The Hall effect, as it has come to be known, was discovered by E.H. Hall in 1879. Hall noted a very small voltage was generated in the transverse direction across a conductor carrying a current in the presence of an external magnetic field (Fig. 19.93), in accordance with the following equation (White, 1988):

$$V_h = \frac{R_h I B}{t}$$

where

V_h = Hall voltage,
R_h = material-dependent Hall coefficient,
I = current in amps,
B = magnetic flux density (perpendicular to I) in Gauss, and
t = element thickness in centimeters.

It was not until the advent of semiconductor technology (heralded by the invention of the transistor in 1948) that this important observation could be put to any practical use. Even so, early silicon implementations were plagued by a number of shortcomings that slowed popular acceptance, including high cost, temperature instabilities, and otherwise poor reliability (McDermott, 1969). Subsequent advances in integrated circuit technology (i.e., monolithic designs, new materials, and internal temperature compensation) have significantly improved both stability and sensitivity. With a 100-mA current flow through indium arsenide (InAs), for example, an output voltage of 60 mV can be generated with a flux density (B) of 10 kG (Hines, 1992). Large-volume applications in the automotive industry (such as distributor timing in electronic ignition systems) helped push the technology into the forefront in the late 1970s (White, 1988). Potential robotic utilization includes position and speed sensing, motor commutation (Manolis, 1993), guidepath following, and magnetic compasses.

The linear relationship of output voltage to transverse magnetic field intensity is an important feature contributing to the popularity of the modern *Hall-effect sensor*. To improve stability, *linear Hall-effect sensors* are generally packaged with an integral voltage regulator and output amplifier. The output voltage V_o fluctuates above and below a zero-field equilibrium position (usually half the power supply voltage V_{cc}), with the magnitude and direction of the offset determined by the field strength and polarity, respectively (White, 1988). (Note also that any deviation in *field direction* away from the perpendicular will also affect the magnitude of the voltage swing.) Frequency responses over 100 kHz are easily achieved (Wood, 1986).

The addition of a *Schmitt-trigger* threshold detector and an appropriate output driver transforms the linear Hall-effect sensor into a digital *Hall-effect switch*. Most commercially available devices employ transistor drivers that provide an open-circuit output in the absence of a magnetic field (Wood, 1986).

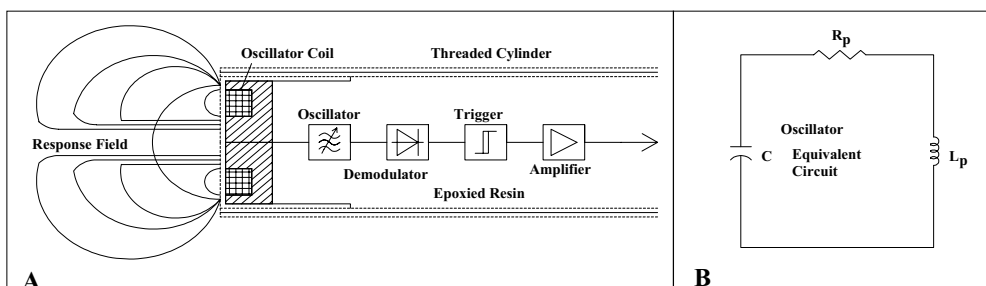



FIGURE 19.94 (A) Block diagram of a typical *ECKO-type* inductive proximity sensor (adapted from Smith, 1985), and (B) equivalent oscillator circuit (adapted from Carr, 1987).

The detector trip point is set to some nominal value above the zero-field equilibrium voltage, and when this threshold is exceeded, the output driver toggles to the *on* state (*source* or *sink*, depending on whether PNP or NPN transistor drivers are employed). A major significance of this design approach is the resulting insensitivity of the Hall-effect switch to reverse magnetic polarity. While the mere approach of the south pole of a permanent magnet will activate the device, even direct contact by the north pole will have no effect on switching action, as the amplified output voltage actually falls further away from the *Schmitt-trigger* setpoint. Switching response times are very rapid, typically in the 400-ns range (Wood, 1986).

Inductive Proximity Sensors

Inductive proximity switches are today the most commonly employed industrial sensors (Moldoveanu, 1993) for detection of ferrous and nonferrous metal objects (i.e., steel, brass, aluminum, copper) over short distances. Cylindrical configurations as small as 4 mm in diameter have been available for over a decade (Smith, 1985). Because of the inherent ability to sense through nonmetallic materials, these sensors can be coated, potted, or otherwise sealed, permitting operation in contaminated work areas, or even submerged in fluids. Frequency responses up to 10 kHz can typically be achieved (Carr, 1987).

Inductive proximity sensors generate an oscillatory RF field (i.e., 100 kHz to 1 MHz) around a coil of wire typically wound around a ferrite core. When a metallic object enters the defined field projecting from the sensor face, eddy currents are induced in the target surface. These eddy currents produce a secondary magnetic field that interacts with field of the probe, thereby loading the probe oscillator. The effective impedance of the probe coil changes, resulting in an oscillator frequency shift (or amplitude change) that is converted into an output signal proportional to the sensed gap between probe and target.

A block diagram of a typical inductive proximity sensor is depicted in Fig. 19.94(A). The oscillator comprises an active device (i.e., a transistor or IC) and the sensor probe coil itself. An equivalent circuit (Fig. 19.94(B)) representing this configuration is presented by Carr (1987), wherein the probe coil is modeled as an inductor L_p with a series resistor R_p, and the connecting cable between the coil and the active element shown as a capacitance C. In the case of a typical Collpitts oscillator, the probe-cable combination is part of a resonant frequency tank circuit.

As a conductive target enters the field, the effects of the resistive component R_p dominate, and resistive losses of the tank circuit increase, loading (i.e., damping) the oscillator (Carr, 1987). As the gap becomes smaller, the amplitude of the oscillator output continues to decrease, until a point is reached where oscillation can no longer be sustained. This effect gives rise to the special nomenclature of an eddy-current-killed oscillator (ECKO) for this type of configuration. Sensing gaps smaller than this minimum threshold (typically from 0.005 to 0.020 in.) are not quantified in terms of an oscillator amplitude that correlates with range, and thus constitute a dead-band region for which no analog output is available.

Monitoring the oscillator output amplitude with an internal threshold detector creates an *inductive proximity switch* with a digital *on/off* output (Fig. 19.95). As the metal target approaches the sensor face, the oscillator output voltage falls off as shown, eventually dropping below a preset *trigger level*, whereupon the threshold comparator toggles from an *off* state to an *on* state. Increasing the gap distance causes the

TABLE 19.4 Nominal Sensing Ranges for Material other than Mild Steel Must be Adjusted Using the Above Attenuation Factors (Smith, 1985)

Material	Attenuation Factor
Cast Iron	1.10
Mild Steel	1.00
Stainless Steel	0.70–0.90
Brass	0.45
Aluminum	0.40
Copper	0.35

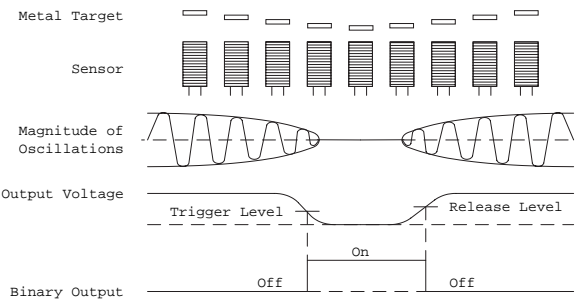



FIGURE 19.95 A small difference between the trigger and release levels (*hysteresis*) eliminates output instability as the target moves in and out of range (adapted from Moldoveanu, 1993).

voltage to again rise, and the output switches *off* as the *release level* is exceeded. The intentional small difference between the trigger level and the release level, termed *hysteresis*, prevents output instabilities near the detection threshold. Typical hysteresis values (in terms of gap distance) range from 3% to 20% of the maximum effective range (Damuck & Perrotti, 1993).

Effective sensing range is approximately equal to the diameter of the sensing coil (Koenigsburg, 1982) and is influenced by target material, size, and shape. The industry standard target (for which the nominal sensing distance is specified) is a 1-mm-thick square of mild steel of the same size as the diameter of the sensor, or three times the nominal sensing distance, whichever is greater (Flueckiger, 1992). For ferrous metals, increased target thickness has a negligible effect (Damuck & Perrotti, 1993). More conductive nonferrous target materials such as copper and aluminum result in reduced detection range, as illustrated in Table 19.4. For such nonferrous metals, greater sensing distances (roughly equivalent to that of steel) can be achieved with thin-foil targets having a thickness less than their internal field attenuation distance (Smith, 1985). This phenomenon is known as the *foil effect* and results from the full RF field penetration setting up additional surface eddy currents on the reverse side of the target (Damuck & Perrotti, 1993).

There are two basic types of inductive proximity sensors: (1) *shielded* (Fig. 19.96(A)) and (2) *unshielded* (Fig. 19.96(B)). If an unshielded device is mounted in a metal surface, the close proximity of the surrounding metal will effectively saturate the sensor and preclude operation altogether (Swanson, 1985). To overcome this problem, the shielded configuration incorporates a coaxial metal ring surrounding the core, thus focusing the field to the front and effectively precluding lateral detection (Flueckiger, 1992). There is an associated penalty in maximum effective range, as shielded sensors can only detect out to about half the distance of an unshielded device of equivalent diameter (Swanson, 1985).

Mutual interference between inductive proximity sensors operating at the same frequency can result if the units are installed with a lateral spacing of less than twice the sensor diameter. This interference typically manifests itself in the form of an unstable pulsing of the output signal, or reduced effective range, and is most likely to occur in the situation where one sensor is undamped and the other is in the hysteresis range (Smith, 1985). Half the recommended 2*d* lateral spacing is generally sufficient for elimination of mutual

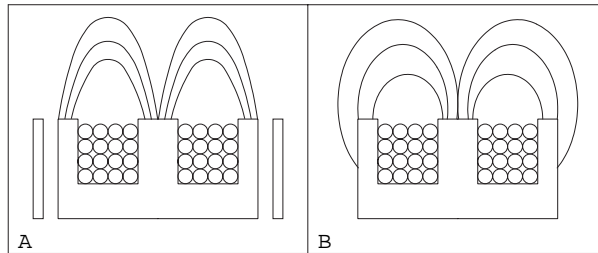



FIGURE 19.96 *Shielded* inductive sensors (A) can be embedded in metal without affecting performance, while the *unshielded* variety (B) must be mounted on nonmetallic surfaces only (Flueckiger, 1992).

interaction in the case of shielded sensors (Gatzios & Ben-Ari, 1986). When mounting in an opposed facing configuration, these minimal separation distances should be doubled.

Capacitive Proximity Sensors

The *capacitive proximity sensor* is very similar to the previously discussed *inductive proximity sensor*, except that the capacitive type can reliably detect dielectric materials in addition to metals. Effective for short-range detection out to a few inches, such sensors react to the variation in electrical capacitance between a probe (or plate) and its surrounding environment. As an object draws near, the changing geometry and/or dielectric characteristics within the sensing region cause the capacitance to increase. This change in capacitance can be sensed in a number of different ways: (1) an increase in current flow through the probe (Hall, 1984), (2) initiation of oscillation in an RC circuit (McMahon, 1987), or (3) a decrease in the frequency of an ongoing oscillation (Vranish et al., 1991). Typical industrial applications include level sensing for various materials (i.e., liquids, pellets, and powders) and product detection, particularly through nonmetallic packaging.

Ultrasonic Proximity Sensors

All of the preceding proximity sensors relied on target presence to directly change some electrical characteristic or property (i.e., inductance, capacitance) associated with the sense circuitry itself. The ultrasonic proximity sensor is an example of a *reflective* sensor that responds to changes in the amount of emitted energy returned to a detector after interaction with the target of interest. Typical systems consist of two transducers (one to transmit and one to receive the returned energy), although the relatively slow speed of sound makes it possible to operate in the transceiver mode with a common transducer. The transmitter emits a longitudinal wave in the ultrasonic region of the acoustical spectrum (typically 20–200 kHz), above the normal limits of human hearing.

Ultrasonic proximity sensors are useful over distances out to several feet for detecting most objects, liquid and solid. If an object enters the acoustical field, energy is reflected back to the receiver. As is the case with any reflective sensor, maximum detection range is dependent not only on emitted power levels, but also on the target cross-sectional area, reflectivity, and directivity. Once the received signal amplitude reaches a preset threshold, the sensor output changes state, indicating detection. Due in part to the advent of low-cost microcontrollers, such devices have for most situations been replaced by more versatile ultrasonic ranging systems that provide a quantitative indicator of distance to the detected object (section "Ultrasonic TOF Systems").

Microwave Proximity Sensors

Microwave proximity sensors operate at distances of 5–150 ft or more (Williams, 1989) and are very similar to the ultrasonic units discussed above, except that electromagnetic energy in the microwave region of the RF energy spectrum is emitted. The FCC has allocated 10.50–10.55 GHz and 24.075–24.175 GHz for microwave field-disturbance sensors of this type (Schultz, 1993). When the presence of a suitable

target reflects sufficient energy from the transmitting antenna back to a separate receiving antenna (see Fig. 19.68 in section "Microwave Range Sensors"), the output changes state to indicate an object is present within the field of view. An alternative configuration employing a single transmit/receive antenna monitors the Doppler shift induced by a moving target to detect relative motion as opposed to presence. These sensors are usually larger than inductive and capacitive sensors, and they are best suited to detect larger objects.

Optical Proximity Sensors

Optical (photoelectric) sensors commonly employed in industrial applications can be broken down into three basic groups: (1) *opposed*, (2) *retroreflective*, and (3) *diffuse*. (The first two of these categories are not really "proximity" sensors in the strictest sense of the terminology.) Effective ranges vary from a few inches out to several hundred feet. Common robotic applications include floor sensing, navigational referencing, and collision avoidance. Industrial applications include sensing presence at a given maximum range (for counting, or to work on a part), sensing intrusion for safety systems, alignment, etc. Modulated near-infrared energy is typically employed to reduce the effects of ambient lighting, thus achieving the required signal-to-noise ratio for reliable operation. Visible-red wavelengths are sometimes used to assist in installation alignment and system diagnostics.

Actual performance depends on several factors. Effective range is a function of the physical characteristics (i.e., size, shape, reflectivity, and material) of the object to be detected, its speed and direction of motion, the design of the sensor, and the quality and quantity of energy it radiates or receives. Repeatability in detection is based on the size of the target object, changes in ambient conditions, variations in reflectivity or other material characteristics of the target, and the stability of the electronic circuitry itself. Unique operational characteristics of each particular type can often be exploited to optimize performance in accordance with the needs of the application.

Opposed Mode

Commonly called an "electric eye" at the time, the first of these categories was introduced into a variety of applications back in the early 1950s, to include parts counters, automatic door openers, annunciators, and security systems. Separate transmitting and receiving elements are physically located on either side of the region of interest; the transmitter emits a beam of light, often supplied in more recent configurations by an LED that is focused onto a photosensitive receiver. Any object passing between the emitter and receiver breaks the beam, disrupting the circuit. Effective ranges of hundreds of feet or more are routinely possible and often employed in security applications.

Retroreflective Mode

Retroreflective sensors evolved from the *opposed* variety through the use of a mirror to reflect the emitted energy back to a detector located directly alongside the transmitter. *Corner-cube retroreflectors* (Fig. 19.97) eventually replaced the mirrors to cut down on critical alignment needs. Corner-cube prisms have three mutually perpendicular reflective surfaces and a hypotenuse face; light entering through the hypotenuse face is reflected by each of the surfaces and returned back through the face to its source. A good retroreflective target will return about 3000 times as much energy to the sensor as would be reflected from a sheet of white typing paper (Banner, 1993). In most factory automation scenarios, the object of interest is detected when it breaks the beam, although some applications call for placing the retroreflector on the item itself.

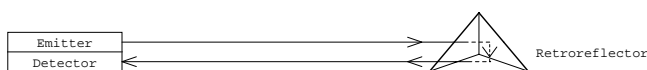



FIGURE 19.97 Corner-cube retroreflectors are employed to increase effective range and simplify alignment (adapted from Banner, 1993).

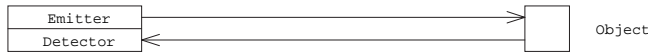



FIGURE 19.98 *Diffuse-mode proximity sensors* rely on energy reflected directly from the target surface.

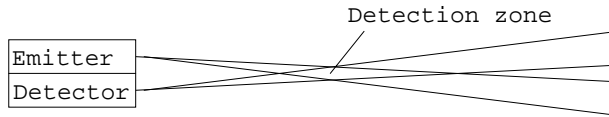



FIGURE 19.99 *Diffuse proximity sensors* configured in the *convergent mode* can be used to ascertain approximate distance to an object.

Diffuse Mode

Optical proximity sensors in the *diffuse* category operate in similar fashion to *retroreflective* types, except that energy is returned from the surface of the object of interest, instead of from a *co-operative reflector* (Fig. 19.98). This feature facilitates random object detection in unstructured environments.

There are several advantages of this type of sensor over ultrasonic ranging for close-proximity object detection. There is no appreciable time lag since optical energy propagates at the speed of light, whereas up to a full second can be required to update a sequentially fired ultrasonic array of only 12 sensors. In addition, optical energy can be easily focused to eliminate adjacent sensor interaction, thereby allowing multiple sensors to be fired simultaneously. Finally, the shorter wavelengths involved greatly reduce problems due to specular reflection, resulting in more effective detection of off-normal surfaces. The disadvantage, of course, is that no direct range measurement is provided, and variations in target reflectivity can sometimes create erratic results. One method for addressing this limitation is discussed in the next section.

Convergent Mode

Diffuse proximity sensors can employ a special geometry in the configuration of the transmitter with respect to the receiver to ensure more precise positioning information. The optical axis of the transmitting LED is angled with respect to that of the detector, so the two intersect only over a narrowly defined region as illustrated in Fig. 19.99. It is only at this specified distance from the device that a target can be in position to reflect energy back to the detector. Consequently, most targets beyond this range are not detected. This feature decouples the proximity sensor from dependence on the reflectivity of the target surface and is useful where targets are not well displaced from background objects.

References

Adams, M.D., "Amplitude modulated optical range data analysis in mobile robotics," *IEEE International Conference on Robotics and Automation,* Atlanta, GA, pp. 8–13, 1993.

Aloimonos, J., Weiss, I., Bandyopadhyay, A., "Active vision," *First International Conference on Computer Vision,* pp. 35–54, 1987.

Arkin, R.C., "Motor-schema-based mobile robot navigation," *International Journal of Robotics Research,* Vol. 8., No. 4, pp. 92–112, Aug., 1989.

Ascension Technologies, P.O. Box 527, Burlington, VT 05402, USA. www.ascension-tech.com.

Ball, D., "Sensor selection guide," *Sensors,* pp. 50–53, April, 1986.

Banner, *Handbook of Photoelectric Sensing,* Banner Engineering Corp., Minneapolis, MN, 1993.

Besl, P.J., "Range imaging sensors," GMR-6090, General Motors Research Laboratory, 1988.

Biber, C., Ellin, S., Shenk, E., "The polaroid ultrasonic ranging system," Audio Engineering Society, 67th Convention, New York, NY, Oct.–Nov., 1980.

Blais, F., Rioux, M., Domey, J., Beraldin, J.A., "A very compact real time 3-D range sensor for mobile robot applications," SPIE Vol. 1007, Mobile Robots III, Cambridge, MA, Nov., 1988.

Borenstein, J., Koren, Y., "Real-time obstacle avoidance for fast mobile robots in cluttered environments," *IEEE International Conference on Robotics and Automation,* Vol. CH2876-1, Cincinnati, OH, pp. 572–577, May, 1990.

Burreson, B., "Magnetic proximity switches in severe environments," *Sensors,* pp. 28–36, June, 1989.

Burt, P.J., Anadan, P., Hanna, K., van der Wal, G., "A front end vision processor for unmanned vehicles," *Advanced Image Processing Group,* David Sarnoff Research Center, Princeton, NJ, April, 1992.

Burt, P.J., Anadan, P., Hanna, K., van der Wal, G., Bassman, R., "A front end vision processor for vehicle navigation," *International Conference on Intelligent Systems,* pp. 653–662, Feb., 1993.

Carr, W.W., "Eddy current proximity sensors," *Sensors,* pp. 23–25, Nov., 1987.

Chen, Y.D., Ni, J., Wu, S.M., "Dynamic calibration and compensation of a 3-D lasar radar scanning system," *IEEE International Conference on Robotics and Automation,* Atlanta, GA, Vol. 3, pp. 652–664, May, 1993.

Damuck, N., Perrotti, J., "Getting the most out of your inductive proximity switch," *Sensors,* pp. 25–27, Aug., 1993.

Depkovich, T., Wolfe, W., "Definition of requirements and components for a robotic locating system," Final Report MCR-83-669, Martin Marietta Denver Aerospace, Denver, CO, Feb., 1984.

Everett, H.R., "A multi-element ultrasonic ranging array," *Robotics Age,* pp. 13–20, July, 1985.

Everett, H.R., DeMuth, D.E., Stitz, E.H., "Survey of collision avoidance and ranging sensors for mobile robots," *Technical Report 1194,* Naval Command Control and Ocean Surveillance Center, San Diego, CA, Dec., 1992.

Figueroa, F., Barbieri, E., "Increased measurement range via frequency division in ultrasonic phase detection methods," *Acustica,* Vol. 73, pp. 47–49, 1991a.

Figueroa, J.F., Barbieri, E., "An ultrasonic ranging system for structural vibration measurements," *IEEE Transactions on Instrumentation and Measurement,* Vol. 40, No. 4, pp. 764–769, Aug., 1991b.

Figueroa, J.F., Lamancusa, J.S., "A method for accurate detection of time of arrival: analysis and design of an ultrasonic ranging system," *Journal of the Acoustical Society of America,* Vol. 91, No. 1, pp. 486–494, Jan., 1992.

Fernando Figueroa and Evangelos Doussis, "A hardware-level method to improve the range and accuracy of an ultrasonic ranging system," *Acustica,* Vol. 78, No. 4, pp. 226–232, May, 1993.

Flueckiger, N., "Inductive proximity sensors: theory and applications," *Sensors,* pp. 11–13, May, 1992.

Gatzios, N.E., Ben-Ari, H., "Proximity control primer," *Sensors,* pp. 47–49, April, 1986.

Gustavson, R.L., Davis, T.E., "Diode-laser radar for low-cost weapon guidance," *SPIE,* Vol. 1633, Laser Radar VII, Los Angeles, CA, pp. 21–32, Jan., 1992.

Hall, D.J., "Robotic sensing devices," Report No. CMU-RI-TR-84-3, Carnegie-Mellon University, Pittsburgh, PA, March, 1984.

Hamlin, "The versatile magnetic proximity sensor," *Sensors,* pp. 16–22, May, 1988.

Hammond, W., "Vehicular use of ultrasonic systems," *Technical Report,* Cybermotion, Salem, VA, May, 1994.

Hebert, M., Krotkov, E., "3-D measurements from imaging laser radars: how good are they?" *International Conference on Intelligent Robots and Systems,* pp. 359–364, 1991.

Hines, R., "Hall effect sensors in Paddlewheel Flowmeters," *Sensors,* pp. 32–33, Jan., 1992.

Jarvis, R.A., "A perspective on range finding techniques for computer vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. PAMI-1, No. 2, pp. 122–139, March, 1983a.

Jarvis, R.A., "A laser time-of-flight range scanner for robotic vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. PAMI-5, No. 5, pp. 505–512, Sep., 1983b.

Kent, E.W., et al., "Real-time cooperative interaction between structured light and reflectance ranging for robot guidance," *Robotica,* Vol. 3, pp. 7–11, Jan.–March, 1985.

Kerr, J.R., "Real time imaging rangefinder for autonomous land vehicles," *SPIE,* Vol. 1007, Mobile Robots III, pp. 349–356, Nov., 1988.

Kilough, S.M., Hamel, W.R., "Sensor capabilities for the HERMIES experimental robot," American Nuclear Society, Third Topical Meeting on Robotics and Remote Systems, Charleston, SC, CONF-890304, Section 4-1, pp. 1–7, March, 1989.

Kim, E.J., "Design of a phased sonar array for a mobile robot," *Bachelor's Thesis,* MIT, Cambridge, MA, May, 1986.
Koenigsburg, W.D., "Noncontact distance sensor technology," GTE Laboratories, 40 Sylvan Rd., Waltham, MA, pp. 519–531, March, 1982.
Lang, S., Korba, L., Wong, A., "Characterizing and modeling a sonar ring," *SPIE Mobile Robots IV,* Philadelphia, PA, pp. 291–304, 1989.
Langer, D., Thorpe, C., "Sonar based outdoor vehicle navigation and collision avoidance," *International Conference on Intelligent Robots and Systems,* IROS'92, Raleigh, NC, July, 1992.
LeMoigue, J., Waxman, A.M., "Projected light grids for short range navigation of autonomous robots," *Proceedings, 7th IEEE Conference on Pattern Recognition,* Montreal, Canada, pp. 203–206, 30 July–2 Aug., 1984.
Lenz, J.E., "A review of magnetic sensors," *Proceedings of the IEEE,* Vol. 78, No. 6, June, 1990.
Lewis, R.A., Johnson, A.R., "A scanning laser rangefinder for a robotic vehicle," *5th International Joint Conference on Artificial Intelligence,* pp. 762–768, 1977.
Loewenstein, D., "Computer vision and ranging systems for a ping pong playing robot," *Robotics Age,* pp. 21–25, Aug., 1984.
Manolis, S., "Resolvers vs. rotary encoders for motor commutation and position feedback," *Sensors,* pp. 29–32, March, 1993.
McDermott, J., "The hall effect: success at 90," *Electronic Design 21*, pp. 38–45, 11 Oct., 1969.
McMahon, V.C., "Solutions from capacitive proximity switches," *Sensors,* pp. 31–33, May, 1987.
Moldoveanu, A., "Inductive proximity sensors: fundamentals and standards," *Sensors,* pp. 11–14, June, 1993.
Moravec, H.P., Elfes, A., "High resolution maps from wide angle sonar," *IEEE International Conference on Robotics and Automation,* St. Louis, MO, pp. 116–121, March, 1985.
NASA, "Fast accurate rangefinder," *NASA Tech Brief,* NPO-13460, Winter, 1977.
Nitzan, D., et al. "The measurement and use of registered reflectance and range data in scene analysis," *Proceedings of IEEE,* Vol. 65, No. 2, pp. 206–220, Feb., 1977.
Nitzan, D., "Assessment of robotic sensors," *Proceedings of 1st International Conference on Robotic Vision and Sensory Controls,* pp. 1–11, 1–3 April, 1981.
Peale, S., "Speed/Motion sensing in challenging environments," *Sensors,* pp. 45–46, Jan., 1992.
Pletta, J.B., Amai, W.A., Klarer, P., Frank, D., Carlson, J., Byrne, R., "The remote security station (RSS) final report," Sandia Report SAND92-1947 for DOE under Contract DE-AC04-76DP00789, Sandia National Laboratories, Albuquerque, NM, Oct., 1992.
Poggio, T., "Vision by man and machine," *Scientific America*, Vol. 250, No. 4, pp. 106–116, April, 1984.
Polaroid, "Polaroid ultrasonic ranging system user's manual," Publication No. P1834B, Polaroid Corporation, Cambridge, MA, Dec., 1981.
Polaroid, "Technical specifications for polaroid electrostatic transducer," 7000-Series Product Specification ITP-64, Polaroid Corporation, Cambridge, MA, June, 1987.
Polaroid, "6500-series sonar ranging module," Product Specifications PID 615077, Polaroid Corporation, Cambridge, MA, 11 Oct., 1990.
Polhemus Incorporated, a Rockwell Collins Company, 40 Hercules Drive, P.O. Box 560, Colchester, VT 05446 (www.polhemus.com).
Schwartz, J.T., "Structured light sensors for 3-D robot vision," Technical Report No. 65, Courant Institute of Mathematical Sciences, New York University, undated.
Scott, M.W., "Range imaging laser radar," US Patent 4,935,616, June 19, 1990.
Siuru, B., "The smart vehicles are here," *Popular Electronics,* Vol. 11, No. 1, pp. 41–45, Jan., 1994.
Smith, J.W., "Design and application of inductive proximity sensors," *Sensors,* pp. 9–14, Nov., 1985.
Swain, M.J., Stricker, M., eds., *Promising Directions in Active Vision,* Report from the National Science Foundation Active Vision Workshop, University of Chicago, IL, 1991.
Swanson, R., "Proximity switch application guide," *Sensors,* pp. 20–28, Nov., 1985.

Vranish, J.M., McConnel, R.L., Mahalingam, S., “Capaciflector collision avoidance sensors for robots,” Product Description, NASA Goddard Space Flight Center, Greenbelt, MD, Feb., 1991.
Vuylsteke, P., Price, C.B., Oosterlinck, A., “Image sensors for real-time 3-D acquisition, part 1,” in *Traditional and Non-Traditional Robotic Sensors,* T.C. Henderson, ed., NATO ASI Series, Vol. F63, Springer-Verlag, pp. 187–210, 1990.
Wavering, A.J., Fiala, J.C., Roberts, K.J., Lumia, R., “TRICLOPS: a high-powered trinocular active vision system,” *IEEE International Conference on Robotics and Automation,* pp. 410–417, 1993.
White, D., “The hall effect sensor: basic principles of operation and application,” *Sensors,* pp. 5–11, May, 1988.
Wildes, R.P., “Direct recovery of 3-D scene geometry from binocular stereo disparity,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 13, No. 8, pp. 761–774, Aug., 1991.
Williams, H., “Proximity sensing with microwave technology,” *Sensors,* pp. 6–15, June, 1989.
Wojcik, S., “Noncontact presence sensors for industrial environments,” *Sensors,* pp. 48–54, Feb., 1994.
Wood, T., “The hall effect sensor,” *Sensors*, pp. 27–36, March, 1986.
Woodbury, N., Brubacher, M., Woodbury, J.R., “Noninvasive tank gauging with frequency-modulated laser ranging,” *Sensors,* pp. 27–31, Sep., 1993.
Young, M.S., Li, Y.C., “A high precision ultrasonic system for vibration measurements,” *Rev. Sci. Instrum.,* Vol. 63, No.11, pp. 5435–5441, Nov., 1992.

19.8 Light Detection, Image, and Vision Systems

Stanley S. Ipson

Introduction

Light detectors span a broad spectrum of complexity. The simplest are single sensors whose output signals are easy to interpret and to interface to other components like microprocessors. In contrast, the image sensors in video and digital cameras, incorporating arrays of up to several million detectors, produce output signals which are complicated to interface and require powerful processors to interpret. Regardless of complexity, the purpose of a light detector is to measure light, and the section “Basic Radiometry” introduces a number of radiometric terms that are employed in the characterization of light, light sources, and detectors. However, manufacturers often specify the performance of their devices using photometric units, which take into account the human visual response to light, and so it is necessary to understand both radiometric and photometric measures of light. Sources of light are briefly discussed in section “Light Sources.” There are several types of light detector in common use and the principles of operation and characteristics of the most widely used, including pyroelectric, photoresistive, photodiode, and phototransistor are summarized in section “Light Detectors.” Vision systems have optical components to form an image and an image sensor to convert the light image into an electrical signal. Image formation is reviewed in section “Image Formation,” before introducing the most widely used detectors, based on charge-coupled device (CCD) technology and complementary metal oxide semiconductor (CMOS) technology, in section “Image Sensors.” The elements required to complete a vision system are discussed briefly in the final section.

Basic Radiometry

Visible light is electromagnetic energy radiated with very short wavelengths in the range between about 400 and 700 nm. At shorter wavelengths, to about 30 nm, is invisible ultraviolet light and at longer wavelengths, up to about 0.3 mm, is invisible infrared radiation. Although electromagnetic radiation displays wave behavior including interference and diffraction, it can also behave like a stream of particles and is emitted and absorbed by matter in discrete amounts of energy called photons. The energy ε of a light

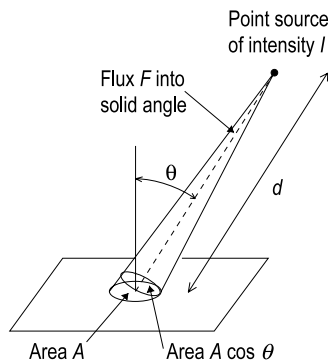



FIGURE 19.100 A point source of intensity I emits radiant power F into the solid angle subtended by the area A. The irradiance at distance d from the source is I/d^2. When the dimensions of A are small compared with d, the solid angle can be approximated by $A\cos(\theta)/d^2$.

photon with wavelength λ is given by

$$\varepsilon = \frac{hc}{\lambda} \tag{19.74}$$

where h is Planck's constant (6.6×10^{-34} J s), c is the speed of light (3×10^8 m s^{-1}) [1]. The most fundamental concept in the measurement of light is radiant power, sometimes called radiant flux (F), which is the flow of energy (photons) per unit time across a specified region in space. It is measured in watts and applies equally to visible and invisible radiation. The corresponding photometric unit is the lumen (lm), which takes into account the varying sensitivity of the eye to light of different wavelengths. One watt of radiation with a wavelength of 555 nm is defined equal to 683 lm. At other wavelengths the number of lumens is reduced (half response at 510 and 610 nm) according to the bell-shaped CIE standard eye-response curve. The remaining radiometric terms, irradiance, intensity, and radiance, are measures of the concentration of light flux. Irradiance (E) is the total radiant power falling on unit area of a surface and is measured in W m^{-2}. The corresponding photometric quantity is illuminance, measured in lm m^{-2}(lux). Radiant intensity (I) is a measure of a point source's ability to illuminate a surface, which decreases as the square of the distance d to the surface. It is measured in Wsr^{-1} and its photometric equivalent is luminous intensity measured in candelas (lm sr^{-1} or cd). The irradiance from a point source of intensity I, falling on a small area A of a surface at distance d with normal inclined at an angle θ to the source as shown in Fig. 19.100, is given by

$$E = \frac{F}{A} = \frac{I\cos\theta}{d^2} \tag{19.75}$$

Although few real sources would seem to be good approximations to a point source (stars in the night sky are exceptions), it is often a good approximation to calculate the irradiance of a surface (detector) by assuming the source has a specified intensity. The error caused by ignoring the spatial extent of the source is less than 1%, if the distance to the source is greater than ten times the largest dimension of the source [2]. When the distance is five (three) times the source size, the error is nearer 4% (9%). Radiant intensity is the most easily measured property of a light source and is often quoted as the performance parameter of a source. Some point sources radiate uniformly in all directions and have an intensity which is independent of direction. Other point sources emit nonuniformly. A Lambertian point source is a source whose intensity varies with direction as $I_0\cos\theta$, where θ is the angle between the measurement direction and the direction of maximum intensity I_0. Many light sources have an intensity which falls off with angle more rapidly than in the case of a Lambertian emitter, which has half intensity at an angle of 60° from the forward direction.

When a source has appreciable spatial extent, its radiance (R) in a given direction is defined as the radiant intensity of the source in that direction divided by the area of the source projected in the same direction, and is measured in W sr^{-1} m^{-2}. Conversely, the intensity of a source is the product of its area

and radiance. Radiance is important in connection with optical systems. In particular, the radiance of the image produced by a lens is equal to that of the object, apart from losses due to absorption and reflection. This fact is used to calculate the image illuminance from an object of specified radiance using a specified lens. The photometric equivalent of radiance is luminance, sometimes loosely called brightness, and is measured in cd m^{-2}. Apart from color, brightness is the only property of light that we can perceive. Because it is normalized by size, the radiance of sources with different sizes can be compared. For example, the radiance of the sun is about 1.3×10^6 W sr^{-1} m^{-2}, the radiance of a 1000 W mercury arc lamp is about 10^7 W sr^{-1} m^{-2}, and the radiance of a 1-mW He–Ne laser with beam diameter of 1 mm and a beam divergence of 1 mrad is about 1.6×10^9 W sr^{-1} m^{-2}. The radiance of most sources increases as the viewing direction approaches the direction normal to the source surface. An extended Lambertian source is an exception, the intensity and projected area vary in the same way with viewing direction, so the radiance is independent of viewing direction.

It is often necessary to estimate the response of a light detector to different light sources. Exact calculations are difficult because the required information may not be available, so it is often better to make a simple estimate and then adjust the equipment to produce the required response. If the source width is small compared with its distance d from the detector then the radiation falling on the detector can be estimated by assuming that the source is a Lambertian point source of intensity I equal to the product of its brightness B and its area S. The irradiance falling on the detector is then given by [2]

$$E = \frac{I\cos\phi}{d^2} = \frac{BS\cos\theta\cos\phi}{d^2} \qquad (19.76)$$

where θ and ϕ are the angles between the normals of the source and detector, respectively, and the line connecting source and detector. If the source is circular, its solid angle S/d^2 can be approximated by $\pi\alpha^2$, providing α the angle in radians equal to the radius of the source divided by its distance from the detector is small. Light detectors have a directional response to radiation, which may be too wide or too narrow for the intended application. The extent of the angular response of a detector to light can be reduced using a collimating tube. Alternatively, if the detector is placed at the focal point of a lens the angular response can be decreased or increased using positive or negative lenses. If the source is effectively a point and all the light brought to focus by the lens falls on the detector, then there is the added advantage of the sensitivity increasing by a factor equal to the ratio of the lens-to-detector area.

Light Sources

The choice of a light detector should take into account the nature of the light source, which might be daylight, a tungsten filament lamp, a quartz halogen lamp, a fluorescent tube, a light emitting diode, etc. The distinct properties of light sources arise partly from their construction and partly from the physical processes which lead to the emission of light [3]. Many sources are thermal in nature; that is, their light emission is due to their high temperature. An object heated to incandescence, such as the filament in a tungsten filament lamp, emits a broad continuous spectrum of electromagnetic radiation, with an intensity that depends on the temperature and its surface emissivity. At any given temperature, no surface emits more radiation than a completely black surface, which has emissivity 1.0 at all wavelengths. It can be important when designing light detecting systems to be aware that the visible light from such lamps (and also fluorescent lamps) is often accompanied by significant amounts of invisible radiation, which detectors may be sensitive to, even if the eye is not. Many light sources operate at temperatures near room temperature and hence are not in thermal equilibrium. Luminescence is the general term used to describe the production of light at a greater rate than that due to the temperature of the body. Common examples of such sources are light-emitting diodes (LEDs), zinc-sulfide electroluminescent panels, and the electron-beam excited phosphors in computer monitor and TV screens. The major properties of sources include: total radiant or luminous power output; efficiency in converting electrical power into radiant power; spectral composition of the output; directionality of the output radiation; area of the

TABLE 19.5 The Characteristics of a Number of Different Types of Light Source

Description	Size	Electrical Input	Light Output	View Angle	Spectral Type
Ultra-bright yellow LED	10 mm dia.	20 mA, 2.1 V	14 cd	4°	Peak at 590 nm
Infrared GaAlAs LED	5 mm dia.	0.1 A, 1.9 V	16 mW sr^{-1}	80°	Peak at 880 nm
Infrared LED	5 mm dia.	0.1 A, 1.9 V	135 mW sr^{-1}	8°	Peak at 880 nm
Small filament lamp	11 mm dia.	6 V, 0.3 A	11 lm	360°	Black body
Miniature fluorescent tube	300 × 16 mm dia.	8 W	480 lm	360°	White
Standard fluorescent tube	1500 × 26 mm dia.	58 W	4800 lm	360°	White
Tungsten halogen dichroic	51 mm dia.	12 V, 20 W	3300 cd	12°	3000 K
Tungsten halogen dichroic	51 mm dia.	12 V, 20 W	460 cd	36°	3000 K

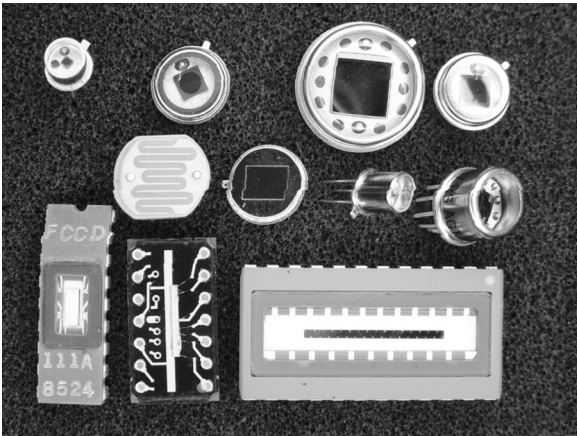



FIGURE 19.101 A collection of light detectors is shown. Along the top row from left to right are four silicon photodiodes with areas of 1, 5, 41.3, and 7.5 mm^2, the last with a photometric color correction filter. Along the middle row from left to right are a CdS photoresistor, a pyroelectric detector, a phototransistor, and a quadrant silicon photodiode containing four separate sensing elements. Along the bottom row from left to right are a 256-element linear CCD, a 64-element charge integrating CMOS array, and a 16-element linear silicon photodiode in a 24-pin d.i.l. package. The diode pitch is 1 mm.

emitting surface; lamp size and operating temperature. Table 19.5 lists characteristics of a number of common types of light sources taken from the lamp suppliers data sheets.

Light Detectors

A light detector converts the radiant power it absorbs into a change of a device parameter such as resistance, surface charge, current, or voltage. A number of light detectors are shown in Fig. 19.101. Some signal conditioning electronics may also be needed to convert the basic output from the detector into a more useful voltage signal, for example, for digitization by an analog-to-digital converter (ADC). This may be integrated into the detector or require external components. Light detectors can be divided into two main types, thermal or photon devices. In thermal detectors, the heating effect of the absorbed radiation results in a change in a temperature dependent parameter, such as electrical resistance (in bolometers) or thermoelectric emf (in thermopiles). The output of thermal detectors is usually proportional to the radiant power absorbed in the detector, and provided the absorption efficiency is the same at all wavelengths, the output is independent of wavelength. The most widely used type of thermal detector is the pyroelectric detector, which is discussed in the next section. Photon detectors, in contrast to thermal detectors, depend on the generation of free charge by the absorption of individual photons. This photon-induced charge causes a change in device resistance, in the case of photoresistors, or an

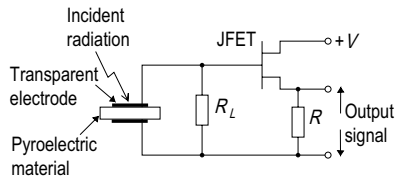



FIGURE 19.102 The basic components within a pyroelectric detector are indicated. An increase in radiation falling on the pyroelectric material causes its temperature to rise and the charge on its surface to change. A transient current flows through the resistor R_L which is of the order of 10^{11} Ω. The JFET reduces the output impedance to R.

output current or output voltage, in the case of photodiodes and transistors. All these photon detectors require a minimum photon energy to create mobile electrons and consequently have a maximum wavelength, dependent on the detector material, beyond which they do not operate. On the other hand, photon detectors generally respond faster to changes in radiation level than thermal detectors and are more sensitive.

Pyroelectric Detectors

Pyroelectric detectors employ a ferroelectric ceramic material (such as lead zirconate or lithium tantalate) which has molecules with a permanent electric dipole moment [4]. Below a critical temperature, known as the Curie temperature, the dipoles are partly aligned and give rise to a net electrical polarization for the whole crystal. As the material is heated and its temperature rises, increased thermal agitation of the molecules reduces the net polarization, which falls to zero at the Curie temperature. The basic detector, shown in Fig. 19.102, consists of a thin slab of ferroelectric material fabricated so that the polarization is normal to the large area faces on which transparent electrodes are evaporated. These are connected together via a load resistor (up to 10^{11} Ω). An increase in radiation falling on the detector makes its temperature rise and causes the captive surface charge, which is proportional to the polarization, to change. This causes a change in the charge induced in the electrodes and a current to flow in the load resistance. Because of the large value of the load resistor used in pyroelectric detectors, an impedance matching circuit, such as a JFET source following circuit, is usually built into the detector as shown in Fig. 19.102. Pyroelectric detectors only respond to changing irradiation and typically can detect radiation powers down to about 10^{-8} W at 1 Hz. Because they respond to the heating caused by absorption of the radiation, they have a wide spectral response. They are useful as low-cost infrared detectors, intruder alarms, and fire detectors.

Photon Detectors

The most widely used photon detectors are made from a semiconducting material. In semiconductors, the electrons fill the available energy levels in the material up to the top of the valence band (VB), which is separated from the bottom of the empty conduction band (CB) by an energy gap E_g, which is characteristic of the material. These energy bands are completely full or empty, respectively, only at a temperature of absolute zero (0 K). At a higher temperature, an equilibrium is reached between the thermal excitation of electrons across the gap (producing free electrons in the CB and positively charged free holes in the VB) and the recombination of pairs of free electrons and holes. The equilibrium number of free electrons and holes increases rapidly with temperature (T) according to the Boltzmann factor $\exp(-E_g/kT)$, where k is Boltzmann's constant (1.38×10^{-23} J/K). This equilibrium is disturbed when photons, with energy greater than E_g, are absorbed by electrons which are excited across the gap. When the radiation source is removed, the number of excess electrons and holes quickly falls back to zero over a time period governed by the recombination time of the material. While excess free charge is present there is a measurable change in the electrical conductivity and this is used in photoresistive (also called photoconductive) detectors. Alternatively, in junction detectors, the rate of generation of photocharge is converted to an output current, or voltage. All semiconductor photon detectors have a relatively narrow

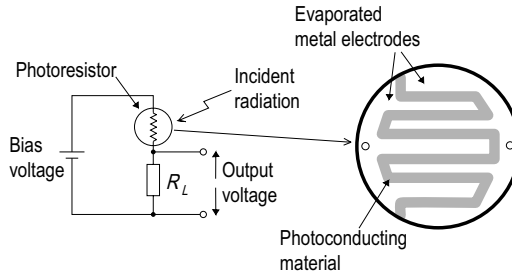



FIGURE 19.103 A simple light detector circuit employing a photoresistor is shown. An increase in light illumination causes the resistance of the photoresistor to decrease and the output voltage to increase. The comb-like pattern typically employed in photoresistors gives a relatively large active area of photoconducting material and a small electrode spacing resulting in high sensitivity.

spectral response, which peaks at a wavelength about hc/E_g. Photoresistors and junction detectors are discussed in more detail in the following sections.

Photoresistors

The electrical conductivity of a semiconductor is the sum of two terms [5], one contributed by electrons and the other by holes, as follows:

$$\sigma = ne\mu_n + pe\mu_p \tag{19.77}$$

Each term is proportional to $n(p)$ the number of electrons (holes) per unit volume in the conduction (valence) band, the electron (hole) mobility $\mu_n(\mu_p)$, and the magnitude of the charge of the electron e. The increase in conductivity, caused by the absorption of photons increasing n and p, is the basis for the operation of the photoresistive detector. This consists of a slab of semiconductor material on the faces of which electrodes are deposited to allow the resistance to be monitored, as illustrated in Fig. 19.103. The photon-induced current is proportional to the length of the electrodes and inversely proportional to their separation, hence the typical comb-like electrode geometry of photoresistors, shown in Fig. 19.73. Because the resistance R_C is inversely proportional to conductivity, the variation of R_C with incident power P_D is very nonlinear and is often expressed in the form

$$\log_{10} R_C = a - b\log P_D \tag{19.78}$$

where a and b are constants. Cadmium sulfide is commonly used as a detector of visible radiation because it is low cost and its response is similar to that of the human eye. Other photoconductive materials include lead sulfide, with a useful response from 1000 to 3400 nm, indium antimonide with a useful response out to 7000 nm, and mercury cadmium telluride with peak sensitivity in the range 5000–14,000 nm. The wavelength range 5000–14,000 nm is of importance because it covers the peak emission from bodies near and above ambient temperature and also corresponds to a region of good transmission through the atmosphere. Photoconductive devices used for the detection of long wavelength infrared radiation should be cooled because of the noise caused by fluctuations in the thermal generation of charge. As a rough rule of thumb, because of the Boltzmann factor, a detector with energy gap E_g should be cooled to a temperature less than $E_g/25k$.

Junction Detectors

In photoresistors, the rate of generation of electron–hole pairs by the absorption of radiation, combined with recombination at a rate characteristic of the device, results in an increase in free charge and therefore electrical conductivity. In junction photodetectors [6], such as photodiodes and phototransistors, newly generated electron–hole pairs separate before they can recombine so that a photon-induced electric

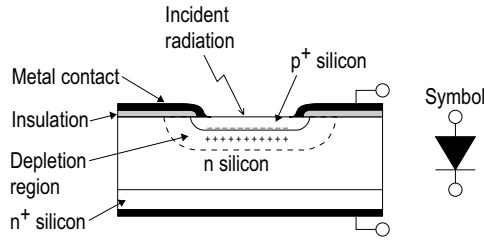



FIGURE 19.104 The basic structure of a typical silicon photodiode is illustrated. A space charge, or depletion region, is formed by the diffusion of mobile charge across the surface between the p-type and n-type silicon. It extends furthest into the n-type silicon because this is more lightly doped than the p-type silicon. Any electron hole pairs generated in this region are prevented from recombining by the presence of the electric field, which sweeps them apart, allowing them to contribute to the photon generated current. The p-type region is made thin to allow photons to penetrate into the depletion region.

current can be detected. The separation of electrons and holes takes place in the electric field associated with a P-N junction fabricated in a semiconductor material, which is usually silicon. The structure of a typical silicon photodiode is shown in Fig. 19.104. The substrate material is lightly doped n-type silicon, which is pure group IV silicon into which has been added a small amount of a group V impurity element. This contributes free electrons to the conduction band of the silicon leaving the impurity atoms ionized and with a positive charge. A region of heavily doped p-type silicon is formed on the top face of the substrate by adding a group III impurity element, by diffusion or ion implantation for example. The group III atoms contribute free holes to the valence band leaving negatively charged impurity ions. The P-N junction is the boundary surface between the p-type and n-type regions on which the opposite impurity concentrations are equal. The mobile electrons and holes diffuse across the boundary from the side where they are in the majority, to the side where they are in the minority. There they recombine leaving a region containing unscreened positive impurity ions on one side of the junction and a region containing unscreened negatively charged impurity ions on the other. The charged region is called the space charge or depletion region, because it is depleted of free charge. The movement of mobile charge continues until the diffusion driving force is balanced by the opposing electric field created in the depletion region by the separation of charge. When equilibrium has been established, the voltage across the depletion region, called the built-in voltage, is about 0.6 V (for silicon). The depletion region extends much further into the n-type silicon than into the p-type silicon for the photodiode shown in Fig. 19.104 because of the very different doping concentrations. The p-type region is made very thin, so that radiation can pass through it, and metallic contacts are made to the p-type and n-type materials. An ohmic contact forms between a metal and heavily doped silicon and to ensure a good ohmic contact to the lightly doped n-type material, an intermediate more heavily doped n-type region is included as shown.

Electron–hole pairs formed in the depletion region when light with wavelength less than hc/E_g is absorbed are separated by the electric field in this region and can be detected in two ways. If the photodiode is left open circuit, a voltage V_p appears across the diode, varying logarithmically with the incident irradiance P_D as follows:

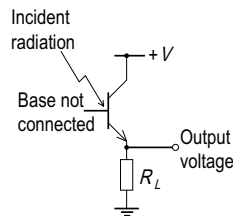
$$V_p = \frac{kT}{e}\ln\left(\frac{\eta P_D A e\lambda}{hci_0}\right) \tag{19.79}$$

where η is the probability of a photon being absorbed, A is the active area of the photodiode, and i_0 is the dark current due to thermal generation. This is the photovoltaic mode of operation. If the diode is operated with a reverse bias, a photon-generated current i_p flows given by the following expression:

$$i_p = \frac{\eta P_D A e\lambda}{hc} \tag{19.80}$$

TABLE 19.6 The Characteristics of a Number of Different Types of Light Detector

Description	Active Region	Response	Spectral Response	Dark Current	Response Time	Acceptance Angle
Medium area silicon photodiode	41.3 mm^2	0.5 A W^{-1} peak	800 nm peak, range 350–1100 nm	4 nA	25 ns	NA
Ultra high speed silicon photodiode	0.5 mm^2	0.35 A W^{-1}	800 nm peak, range 400–1000 nm	10 nA	1 ns	NA
Filtered silicon photodiode	7.5 mm^2	7 nA lux^{-1}	560 nm peak, range 460–750 nm	2 nA	3.5 μs	100°
16 photodiode array on 1 mm pitch	Each diode 0.66 mm^2	0.6 A W^{-1}	900 nm peak, range 400–1100 nm	0.1 nA	4 ns	NA
Silicon phototransistor	0.7 mm^2	9 μA lux^{-1}	880 nm peak, range 450–1100 nm	0.3 μA	15 μs	30°
Silicon phototransistor	0.7 mm^2	2 μA lux^{-1}	880 nm peak, range 450–1100 nm	0.3 μA	15 μs	80°
CdS photoconductor	6.3 mm dia.	9 kΩ at 10 lux, 400 Ω at 1000 lux	530 nm peak	NA	100 ms	NA

FIGURE 19.105 A simple phototransistor light detector circuit is shown. Photon-generated current flowing in the base-collector diode may be amplified several hundred times by transistor action. Although the photon-generated current is much larger than in an equivalent photodiode, response time of the phototransistor is much longer.

In this photoconductive mode, the current through the photodiode varies linearly with light irradiance. The dark current i_0 varies rapidly with temperature and limits the sensitivity of the device but the photoconductive mode generally has faster response, better stability, and wider dynamic range than the photovoltaic mode. The responsivity K_D of the detector is defined by the relation $i_p = K_D P_D$ and is less than 1 A W^{-1} for a silicon diode. In the ideal case, K_D varies linearly with wavelength, according to Eq. (19.80), up to the threshold value set by the energy gap. Photodiodes are available with a wide variety of characteristics differing in sensitivity (area), speed of response, spectral response, and acceptance angle. They are available with single devices or multiple devices (quad, linear array) in a single package.

The output signals from photodiodes needs amplification for many applications. This may be provided by a separate amplifier or by providing internal gain as in the phototransistor. This is constructed so that radiation can fall on the base region of the transistor and the resulting base current is then internally amplified. Often there is no external connection to the base and the amplified photocurrent is monitored using the simple circuit shown in Fig. 19.105. A typical phototransistor has a responsivity several hundred times higher than that of a photodiode but the frequency response is relatively poor. Phototransistors are often integrated with a spectrally matched LED into a single sensor package to act as a proximity sensor, as in end-of-tape sensors, coin detectors, and level sensors. For reference, the characteristics of several different types of discrete light detector are listed in Table 19.6.

By fabricating many small light detectors in a closely spaced array, it is possible to measure light intensity at an array of points over a region. This is ideal for electronic imaging applications involving video and still cameras. Image sensors designed for this purpose are discussed in the section titled "Image Sensors," but first it is useful to consider the formation of the images which the detectors sense.

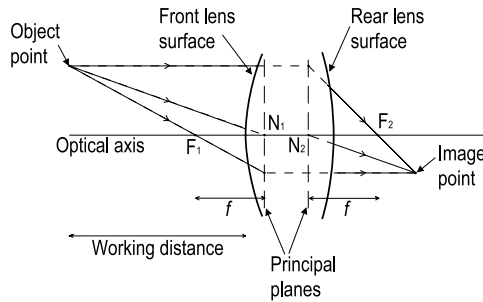



FIGURE 19.106 The cardinal points of a multi-element lens operating in a single medium (usually air) are indicated. The principal points and nodal points then coincide at N_1 and N_2 and the front and rear focal lengths are equal (f). Three rays from an object point are traced through the lens to the corresponding image point using the properties of the cardinal points. In the case shown the image magnification is 0.5, so the image is some distance behind the rear focal point F_2. For distant objects the image plane would coincide with the plane transverse to the optical axis passing through F_2. Lenses are normally corrected for aberrations assuming that the object distance will be greater than the image distance. In this case, for close-up work when the image distance is greater than the object distance, the image quality is improved by reversing the lens.

Image Formation

Although perfect images are formed by small pinholes, lenses are needed to form bright images and range from simple single-component lenses used to increase the amount of light falling on a single detector or in low-cost cameras to complex zoom lenses, with between 14 and 20 components, capable of producing high quality images of varying size. The two most important properties of a lens are its focal length f, which determines the imaging behavior, and its light-gathering power or speed, specified by an f-number $f_{\#}$. A lens has an optical axis passing through the central axis of each of its components along which a ray of light passes without deviation. A lens is characterized, regardless of its complexity, by six cardinal points [2] spaced along the optical axis as illustrated in Fig. 19.106, for a positive converging lens. The position and magnification of the image of an object can be determined using these cardinal points, which include two focal points, two nodal points, and two principal points. The nodal points have the property that a ray outside the lens travelling towards one nodal point emerges from the lens in a parallel direction, appearing to come from the other nodal point. The focal point is the point which a ray of light incident on the lens parallel and close to the optical axis converges to (positive lens) or diverges from (negative lens) after passing through the lens. The point where the lines colinear with the ray on the two sides of the lens intersect defines a point on the principal plane. The point where the optical axis intersects the principal plane is called the principal point. There are two nodal points, focal points and principal planes, because light can be incident on the lens from either side. The front and back focal lengths of the lens are the distances between the front and back focal points and their corresponding principal planes. In the normal situation, when the lens is operating in a single medium, such as air, the positions of the nodal points and principal points coincide and the front and back focal lengths are equal. In general, when the lens construction is asymmetric, the front and back focal points are at different distances from the corresponding external lens surface. In the case of an ideal thin lens, the principal planes coincide with the lens center but in multi-element lenses they may be separated by +20 to −10 mm, depending on the lens design. A lens of focal length f produces an image in best focus at a distance v when the object is at distance u where

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f} \qquad (19.81)$$

and the distances are measured to the corresponding principal planes. The image magnification m, defined as the ratio of image to object sizes, is equal to the ratio of the image to object distances and is related to

the total distance between object and image D_T by

$$D_T = \frac{f(m+1)^2}{m} + D_N \tag{19.82}$$

where D_N is the separation between the principal planes. Lenses generally have a focusing range around 5–10% of the focal length giving a maximum magnification of 0.05–0.1. For larger magnifications extension rings can be fitted between the lens and sensor mounting. The extension size required, with the lens focused at infinity, is simply the product of the magnification and the focal length, since image distance is $f(m + 1)$. When a lens is focused to produce an optimally sharp image for a particular object distance u, there is range of closer and further object distances over which the image is still acceptably sharp. This range is called the depth-of-field or depth-of-view F_o and its size depends on the f-number of the lens, the magnification, and the acceptable blur spot size C in the image plane [7]. The blur spot size depends on the image sensor and for 35 mm film C is usually assumed to be between 0.02 and 0.033 mm, while for an image sensor array C is the separation between the individual detector elements, typically about 0.01 mm. Depth-of-field decreases with magnification and for m greater than 0.1 is calculated using the following formula:

$$F_o = 2f_{\#}\frac{C(m+1)}{m^2} \tag{19.83}$$

The accuracy of alignment required of the image sensor depends on the depth-of-focus F_i, which is the longitudinal range of image positions over which the image is acceptably sharp. Sensor alignment is most critical when the lens f-number is small and the image magnification is also small. When m is small F_i reduces to $2Cf_{\#}\, v/f$ and equals m^2F_o. When m is large, the depth-of-focus is not so critical.

It is frequently necessary to relate the lighting of a scene to the image irradiation falling on a sensor. As accurate calculations are difficult, it is usually best to make a simple estimate and then make fine adjustments to the lighting or lens aperture. When the object of interest in the scene is not a light source but is visible because it is reflecting light, then its luminance must be estimated from the radiation falling on it and the reflection coefficient R_o of its surface [3]. For example, if the object is a Lambertian surface, with illumination L_o, then its luminance is given by

$$B = L_o\frac{R_o}{\pi} \tag{19.84}$$

When a lens is used to form an image of the object, the illuminance on the optical axis in the image plane L_S in lux is related to the luminance of the object B in cd m^{-2} by

$$L_s = \frac{TB\pi}{[2f_{\#}(m+1)]^2} \tag{19.85}$$

where $f_{\#}$ is defined as the ratio of the focal length to diameter of the effective lens aperture and losses in the lens are characterized by a transmission coefficient T[7]. Due to a number of geometrical factors, the image illumination falls off with angle θ from the optic axis as $\cos^4\theta$. Near the axis, the illuminance varies only slowly with angle but at an angle of 30° it has fallen by 44%. Equation (19.85) is the basis for rating the speed of lenses by their f-numbers and indicates that the smaller the f-number, the greater the image illuminance. This formula is appropriate when the magnification is small, but for close-up work, when the image distance is significantly greater than the focal length, the f-number $f_{\#}$ should be replaced by $(m + 1)f_{\#}$ when calculating image illuminance.

Lenses are manufactured to match standard image sizes such as the 36 mm × 24 mm 35 mm photographic format and the standard television sensor sizes 1″, 2/3″, 1/2″, 1/3″, and 1/4″. These sizes are defined to be twice the horizontal dimension of a rectangular image with 4:3 aspect ratio so that, for example, a 1″ sensor has a width of 12.7 mm, a height of 9.5 mm, and a diagonal length of 15.9 mm. Lens sizes are similarly specified to allow easy matching of lenses to sensors. The maximum angular field-of-view F_{OV} of a lens focused at infinity is given by

$$F_{OV} = 2\tan^{-1}\left(\frac{C_F}{2f}\right) \tag{19.86}$$

where C_F is the diagonal of the sensor format. For example, a 35-mm lens with a focal length of 55 mm has a field of view of 43°. Because image distortion and sharpness worsen towards the edges of the field of view it is acceptable, for example, to use a 2/3″ lens with a 1/2″ sensor, but not the converse. A 35-mm camera lens generally performs much better, at relatively low cost, than a corresponding C-mount lens supplied for a TV camera but a C-mount to Pentax, Cannon, or Nikon mount converter is then required.

Image Sensors

The generation of an output signal from a standard image sensor involves up to four operations: the conversion of the spatial distribution of light irradiance in the image plane into a corresponding spatial distribution of charge; the accumulation and storage of this charge near the point where it is generated; the transfer or read-out of this charge; and the conversion of the charge to an output voltage signal. Each of these operations can be achieved in many ways. Vacuum tube sensors such as vidicons, for example, use a photoconductive detector material and a scanning electron beam for read-out while CMOS sensors use a photodiode detector and a readout bus. Solid state devices are currently the most widely used types, so only CCD and CMOS sensors are considered here. In these devices the image irradiance is measured on a one- or two-dimensional array of sample regions with positions fixed during fabrication. Each sample is called a picture element or pixel and the greater the number of pixels, the higher the resolution with which the image can be recorded. Area sensors are manufactured with numbers of pixels ranging from tens of thousands to several million. Color sensors are achieved by placing color filters over the individual pixels, in a mosaic or stripe pattern, and interpolating the color values at pixels where necessary from the neighboring values. In such color devices the color resolution is lower than the luminance resolution, but this is not important for many applications because the resolution of the human eye is worse for color than for luminance. More expensive color cameras use three precisely aligned sensors, one for each primary color. Cameras incorporating such sensors generally produce either a television standard signal [8] (RS-170 monochrome and NTSC color for 525 American television or CCIR monochrome and PAL color for European television) or a digital signal such as RS-423, USB or IEE 1394 Fire Wire, which can be readily connected to a computer.

Charge-Coupled Devices

In a charge-coupled device [9] an isolated packet of charge, of between 10 and 10^6 electrons, is moved through the semiconductor, from a position in one CCD cell to a position in an adjacent cell, by applying a sequence of voltage pulses to gate electrodes. In CCD-based light sensors, photon generated charge packets accumulate in photosites, which are modified CCD cells, and are then transported through other CCD cells to another modified cell with a readout amplifier attached. A CCD is fabricated on a single crystal wafer of P-type silicon and consists of a one- or two-dimensional array of charge storage cells, on centers typically about 10 μm apart. The operation of a 3-phase CCD cell is illustrated in Fig. 19.107. Each cell has three closely spaced electrodes (gates) on top, insulated from the silicon by a thin layer of silicon dioxide. A positive voltage applied to one of these gates will attract and store any free charge generated in the silicon due to light or thermal action while free holes are repelled and collected by the substrate electrode. Lower voltages on the adjacent gates isolate it from the neighboring cells, creating a

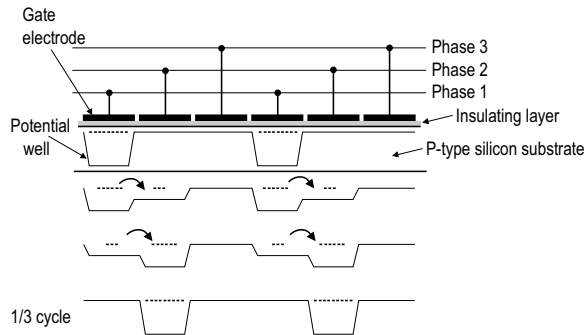



FIGURE 19.107 The movement of charge from one potential well to the next in a 3-phase CCD is illustrated. Each CCD cell has three gate electrodes. In the upper potential diagram, the well is formed under the first electrode in each CCD cell by voltage applied to the phase 1 line. As the voltage is reduced on the phase 1 line and increased on the phase 2 line, the original potential wells collapse and new ones form under the second gate in each cell, causing any charge present in the wells to move sideways as indicated. Two more cycles are required to complete the movement of charge into the first well of the next CCD cell.

localized potential well within the cell. A cell of size 8 μm $\times$ 8 μm can hold about 200,000 electrons before saturating. Cells designed to be light sensitive have electrodes made of semitransparent polysilicon so that light can penetrate into the storage region, while cells intended only for charge transport are covered with a surface layer opaque to light. During operation, the voltages on the electrodes are held constant for a time (integration time) to allow packets of charge to accumulate on the photosites in proportion to the local irradiance. At the end of this time a sequence of voltage pulses are applied to the electrodes to transfer the packets of charge from one storage cell to the next until they reach a sensing amplifier, which generates a voltage which is about 0.6 μV per electron. The charge transfer efficiency (CTE) of real devices is less than 100% and between 99.95% and 99.999% of the stored charge is moved to the next cell, depending on the precise construction and clocking frequency. This allows devices to be manufactured with a line of many hundreds or thousands of storage cells feeding a single amplifier.

Although there are many variations in CCD construction, the basic characteristics of CCDs from different manufactures are similar. CCD light sensors have an inherently linear variation of output voltage with irradiance, from the minimum useful level set by noise to the maximum useful level set by saturation of the output amplifier or by the limited capacity of the charge storage cells. The dynamic range of the device is defined as the ratio of the maximum output signal to the output resulting from noise. Manufacturers sometimes quote noise figures as peak-to-peak or as root-mean-square values (typically five times smaller), but the former value is more relevant for imaging applications. Due to manufacturing limitations, the photosites do not have identical sensitivity and dark signal characteristics. For example, photoresponse nonuniformity (PRNU) is easy to measure using uniform sensor illumination and is typically 5–10%. Its effects can be removed, if necessary, by calibration. The basic spectral response of a silicon sensor extends from 200 to 1100 nm, with a maximum sensitivity of about 1 μA of generated charge per microwatt of incident radiation, but this is modified by the electrode structures formed on the surface of the silicon. Longer wavelength photons penetrate more deeply than shorter wavelength photons and the short wavelength response is typically worsened to 450 nm by absorption in the surface layers. Infrared photons may generate electrons some distance from the point of entry into the silicon, with the result that the charge may be collected by a different cell. This reduces the resolution of the device and if infrared operation is not required, but the illumination contains infrared (for example, from a tungsten lamp), an infrared reflecting filter (a hot-mirror filter) is often used. If the widest possible spectral response is required, devices have the substrate thinned and are operated with the illumination falling on the back surface, which is free of electrodes. Back illuminated devices are fragile and costly but are used in specialist low-light applications like astronomy and biology.

All CCD cells accumulate charge linearly with time due to thermally generated electrons produced within the cells and at electrode interfaces. Like the photoresponse, this dark signal varies from cell to cell and can be compensated for by calibration. These thermally generated contributions are most significant for low-light level applications and can be reduced by cooling the sensor using either a thermoelectric cooler, a Joule Thomson cooler, or a liquid nitrogen dewar. The dark signal reduces by 50% for every 7°C reduction in temperature and at –60°C, produced by a Peltier cooler, the dark signal is typically reduced to about one electron per pixel per second. Another important temperature dependent characteristic of the CCD sensor, which improves with cooling, is the noise floor of the output amplifier which is proportional to $T^{1/2}$ and typically equivalent to about 300 electrons at room temperature. A CCD device used in astronomy illustrates the performance achieved by cooling. Operated at about –110°C, this device has a readout noise of about 10 electrons, a dark current less than 0.3 electrons per minute, and a quantum efficiency for converting visible photons into electrons of between 70% and 80%. Light may be integrated for periods of hours compared with the approximately 1/8 s to 1/4 s integration period of the dark adapted eye. Compared with photographic film previously used for low-light level imaging in astronomy, cooled CCDs are from 10 to 100 times more sensitive, linear in response rather than nonlinear, and have a much greater dynamic range so that both faint and bright objects can be recorded in the same exposure.

The transfer of charge from one cell to the next takes time and the CTE worsens with increasing clocking speed and with cooler temperatures. This limits the number of cells which can be used to transport charge from a photosite to the readout amplifier. It also limits the rate at which data can be transferred out of the CCD and the resulting image transfer rate. However, there are many variations in CCD technology aiming to improve performance. For example, virtual-phase CCDs [10] have some of the electrodes replaced by ion-implanted regions resulting in improved blue response and higher sensitivity, because of the removal of some of the blocking surface gates and simpler drive circuitry due to the lower number of gates per cell. The biggest contribution to the dark signal is defects at interfaces and a manufacturing technique known as pinning can be used to passivate the interface states, producing an order of magnitude improvement in a dark signal as well as improved quantum efficiency and CTE. The readout noise performance can be improved by a signal-processing technique called correlated double sampling. This involves taking the output as the difference between two signals, one with the charge signal present and one without, so that major noise components are cancelled. A number of architectures are employed in CCD devices [11]. Several of these, including linear devices and area devices of the full-frame, frame transfer, and interline transfer types, are discussed in the following sections.

Linear Charge-Coupled Devices

A linear CCD sensor consists of a line of up to several thousand photosites and an adjacent parallel CCD shift register terminated by a sensing amplifier. Each photosite is separated from a shift register cell by a transfer gate. During operation a voltage is applied to each photosite gate to create empty storage wells, which then accumulate amounts of charge proportional to the integral of the light intensity over time. A transfer pulse at the end of the integration period causes all the accumulated charge packets to be transferred through the transfer gates to the shift register cells. The charges are clocked through the shift register to the sensing amplifier producing a sequence of voltage pulses with amplitudes proportional to the integrated light falling on the photosites. In practice it is common for shift registers to be placed on both sides of the photosites with alternate photosites connected by transfer gates to the right and left registers. These halve the time required to clock out all the data. There is a limit to the number of electrons (typically 1000–2000 times the area of the photosite in μm^2) which can be stored in a cell, before electrons start to spill over into adjacent cells. This blooming effect is a problem with images containing intense highlights. It is reduced by about a factor of 100 by adding antiblooming gates between adjacent photosites and transfer gates and channel stops between adjacent photosites. The voltage on the antiblooming gates is set at a value which allows surplus charge to drain away instead of entering the transfer gates and shift register. By clocking this voltage, variable integration times which are less than the frame pulse to frame pulse exposure time can also be attained.

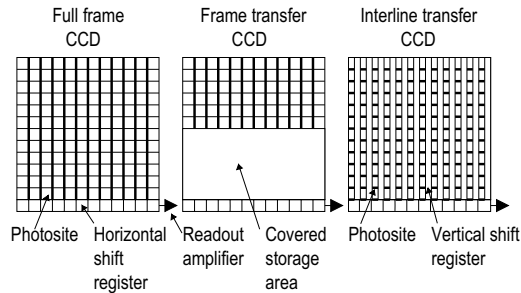



FIGURE 19.108 The three basic architectures used in area CCDs are illustrated. In the full-frame transfer CCD most of the device area is employed as photosites. Photon-generated charge is transferred down each column one cell at a time into the horizontal shift register where it must all be transferred to the readout amplifier before another vertical movement of charge can take place. The frame-transfer CCD reduces the need for a mechanical shutter to prevent charge smearing, which would otherwise occur, by providing a covered storage area into which all the photon-generated charge can be rapidly shifted vertically at the end of the integration period. The interline-transfer CCD allows all the photon-generated charge to be transferred to the covered vertical shift registers in one step, virtually eliminating this source of charge smearing.

Area Charge-Coupled Devices

Three basic architectures are used in area CCDs and are illustrated in Fig. 19.108. The full-frame CCD consists of an imaging area separated from a horizontal CCD shift register by a transfer gate. In the imaging area each photosite is one stage of a vertical shift register separated from neighboring shift registers by channel stops and anti-blooming structures. During the light integration period, the vertical clocks are stopped and the photosites collect photoelectrons. At the end of this period the charge is clocked out vertically, one row at a time into the horizontal shift register. The charge in the horizontal shift register is then very rapidly shifted towards the output amplifier by the application of a horizontal clock signal. For example, the RA1001J, 1024 × 1024 pixel full-frame CCD from EG&G Reticon achieves a readout rate of 30 frames per second. To avoid image smear during the readout period, full-frame sensors must be operated with external shutters or used in low-light level applications requiring very long integration times compared with the readout time, as in astronomy.

The frame-transfer CCD greatly reduces the need for an external shutter by providing a light-shielded storage section into which the entire image charge is shifted at a rate limited primarily by CTE considerations. The charge is read from the storage region during the next integration period without any further image smearing. In some devices, such as the EG&G Reticon RA1102, the storage area is split into two on opposite sides of the imaging area. This improves performance by halving the maximum number of transfers required to reach the nearest storage region. With sensors designed for interlaced operation, as opposed to the non-interlaced progressive scan readout mode, this reduction occurs automatically. Each integration period then corresponds to one video field and only half the number of rows in the frame is required at any one time. For example, to produce an interlaced video frame containing 576 image lines (CCIR standard), a frame transfer sensor with only 288 rows of storage is required. By changing the clock signals, the odd field can be displaced vertically by half a line width relative to the even field. This ensures that the odd and even lines contain different information and reduces aliasing because the cell width is twice the separation between the lines in the frame. Many companies produce frame-transfer CCD sensors and cameras including Cohu, Dalsa, EG&G Reticon, EEV, Kodak, Philips, and Thomson-CSF.

The interline-transfer (ILT) architecture virtually eliminates image smear by providing each column of photosites with an adjacent light-shielded vertical CCD shift register into which the charge is transferred by a transfer pulse. The contents of all the vertical shift registers are then shifted simultaneously one pixel at a time into a horizontal shift register where they are rapidly shifted to an output amplifier. This approach makes it easy to achieve short integration times and true “stop-motion” exposure control with progressive scan. It also increases the “dead space” between the active pixels reducing the sensitivity of

the image sensing area and increasing aliasing effects compared with frame-transfer sensors. For the latter, the fill factor, which is the percentage of the imaging area which is light sensitive, can be close to 100% whereas it is usually less than 50% for interline-transfer devices. Localized bright objects tend to produce vertical streaks in an ILT device because strong light can leak under the narrow light shield covering the vertical shift registers, causing image smearing similar to that in a full frame device. For interlaced operation, two adjacent pixels, for example, 1 and 2, 3 and 4, etc. are transferred to a single shift register cell on one field and in the next field pixels 2 and 3, 4 and 5, etc. are transferred together. This is rather similar to the interlaced operation of a frame transfer CCD. Many companies manufacture ILT CCD sensors and cameras including, Hitachi, NEC, Panasonic, Pulnix, and Sony.

CMOS Sensors

CMOS image sensors are based on a technology that is older than CCD technology [12]. However, CCD sensors originally offered better image quality than CMOS devices could match so they came to dominate the market. There is now renewed interest in the older technology because it potentially offers major advantages over CCDs. The CMOS process used in sensors is similar to that which has been highly developed in order to manufacture dynamic RAM and consequently should be able to produce cheap, small high-resolution, randomly addressed, low-power sensors. It is also possible to integrate image sensing, control, processing, and interfacing on the same chip, so that a camera on a chip is possible using CMOS technology, but not with CCD technology. As a result of recent research and development, several manufacturers are now claiming to have achieved CMOS sensors providing similar quality to that of mainstream CCDs. Manufacturers producing CMOS sensors include Fillfactory, National Semiconductor, Philips, ST Microelectronics, and Y Media.

A CMOS sensor consists of an array of photodiodes, which are connected to readout amplifiers by bus lines and MOS switches. The principle of readout is illustrated in Fig. 19.109. Each pixel is connected to an output amplifier by a switch whose control line is connected to a digital shift register. Shifting a bit through the register connects the photodiodes sequentially to the output amplifier. Random readout of the photodiodes can be achieved by replacing the shift register by an address decoder connected to an address bus. Two-dimensional arrays of photodiodes are connected in a configuration similar to a cross-point switching matrix with a switch and diode at each cross point and separate vertical and horizontal shift registers. To scan the array, the vertical shift register turns on a complete row of switches and the photodiodes in that row output their signals into vertical bus lines. These are connected, in turn, to an

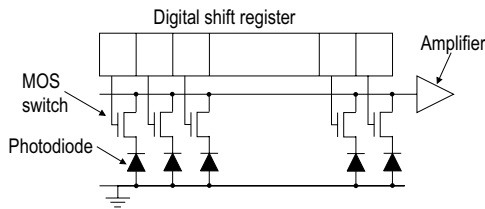
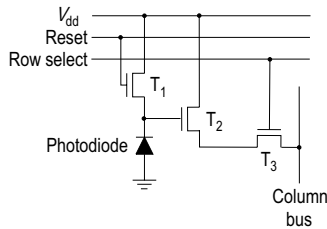



FIGURE 19.109 The principle of readout in a one-dimensional CMOS sensor is illustrated. Each pixel is connected to an output amplifier by a MOS switch whose control line is connected to the digital shift register. Shifting a bit through the register connects the photodiodes, in turn, to the output amplifier.

FIGURE 19.110 A typical CMOS three-transistor active pixel is shown. Transistor T_1 is connected to the reset line allowing the capacitance of the photodiode to be reset at the start of photo-current integration. In the continuous mode CMOS device this transistor is connected to act as high value resistor. Transistors T_2 and T_3 allow the signal to be transferred from the photodiode to the column amplifier, via the column bus line.

output amplifier by a set of switches connected to the horizontal shitch register. Using two separate vertical shift registers and separating row select and row reset functions enables a rolling curtain type of electronic shutter to be implemented with an exposure ratio equal to the number of rows.

Photodiode arrays generally have less extensive electrode structures over each sensing element compared with CCD arrays and consequently the spectral response is smoother and extends further at the blue end of the spectrum, which is an advantage for color sensors. The peak quantum efficiency is also higher ranging from 60% to 80% compared with 10% to 60% for photogates, leading to almost twice the electrical output power for a given light input power. On the other hand, photodiodes have higher noise levels than CCDs because of the reverse-bias leakage current. The photodiode can be operated in integrating mode or in continuous mode. In the former, the photodiode capacitance is reset to a reference reverse bias and then allowed to float. The charge on the photodiode capacitance is then discharged by photon-generated current and leakage currents. After a specified integration time, the remaining charge can be read and the difference from the reference value is proportional to the diode irradiance if the leakage sources are negligible. Both passive and active pixel devices have been developed. In the latter, the charge on the photodiode is read out through a MOS field effect transistor (MOSFET), which converts charge to voltage and provides gain. Figure 19.110 illustrates a typical three-transistor active pixel. Transistor T1 resets the photodiode and after an integration period the pixel signal is read out through the source-follower transistor T2 and the selection transistor T3. Two disadvantages of this approach are a low fill-factor, because of the pixel area used by the amplifying transistor, and the increased pixel nonuniformity because of variations in transistor characteristics. In the ibis range of CMOS sensors designed by FillFactory, a small area photodiode is used which reduces dark current and kTC noise while also increasing the charge-to-voltage conversion factor (9 μV per electron at output in ibis1). The pixel architecture introduces a small potential barrier, which prevents photon-induced electrons from being collected by structures in the pixel other than the photodiode. This allows the photodiode to collect most of the electrons generated in the substrate beneath the pixel effectively increasing the fill-factor [13]. In integrating devices, the fixed pattern noise resulting from variations in MOSFET thresholds can be reduced by correlation double sampling. The pixel is sampled at the end of the integration period and the pixel is reset and sampled again. The difference in the two samples is the measure of the light intensity and is free of pixel offsets. Photoresponse nonuniformity is harder to control. It is caused by variations in the photodiode collection volume, junction capacitance, and gate capacitance of the MOSFET amplifier. In the ibis4 1280 × 1024 pixel sensor, PRNU is quoted as less than 10% peak-to-peak with half saturation in the neighborhood.

A light sensor with a continuous pixel response has the advantage that pixels can be accessed in any order without an integration time so an image processing algorithm could decide, on-the-fly, which pixel or group of pixels to read next. In the Fuga range of sensors designed by FillFactory, continuous mode operation is achieved by passing the photon-generated current through a series resistance [14]. Because the photon-induced current is very small, the series resistance must be very large and it is realized by a MOSFET operated in weak inversion. The resulting current-to-voltage conversion is logarithmic and the devices have a very wide dynamic range (six orders of magnitude or 120 dB) with a quoted dark limit of 10^{-4} W m^{-2}. The output signal of an individual pixel cannot respond instantaneously to a change in irradiance because of the RC time constant of the photodiode capacitance and the series resistance. A typical value for this time constant is a fraction of a millisecond. The fixed pattern noise of these devices due to pixel nonuniformity is very large, around 50–100%, and cannot be reduced by correlated double sampling because of the continuous response. However, it is static and can be greatly reduced using correction values stored in a look-up table implemented in PROM, so that fully corrected monochrome or color cameras with 1024 × 1024 pixels are available with Fuga sensors.

Vision Systems

Machine vision is used in a wide variety of applications including manufacturing operations, measurements in science and engineering, remote surveillance, and robotic guidance. In machine vision systems the principal imaging component is not just a sensor chip, but a complete camera, like those shown in

FIGURE 19.111 Three solid state cameras are shown. The nearest is an inexpensive single board camera with a CMOS sensor and 4-mm lens. The middle one is a miniature CCIR interline frame-transfer CCD camera, dwarfed by the 9.5- to 75-mm C-mount zoom lens. The rear camera is a high quality Cohu 4712 monochrome frame-transfer CCD camera fitted with a 16-mm C-mount lens and 20-mm extension tube.

Fig. 19.111, including sensing array, associated electronics, output signal format, and lens. Depending on the application the camera could be RS-170/CCIR monochrome, NTSC/PAL color, progressive scan, variable scan, or line scan. Five major system parameters which govern the choice of camera are field of view, resolution, working distance, depth of field, and image data acquisition rate. Color may also be important to the application, but otherwise monochrome images are preferred because they require less memory and process faster. As a rule of thumb, for size measurement applications, the sensor should have a number of pixels at least equal to twice the ratio of the largest to smallest object sizes of interest. Lighting should be arranged to illuminate the objects of interest so that the best possible images can be acquired. Lighting might be ambient, high-frequency fluorescent, LED, incandescent, or quartz halogen.

A frame grabber or video capture card, usually in the form of a plug-in board which is installed in the computer, is often required to interface the camera to a host computer. Camera suppliers can recommend compatible frame grabbers. The frame grabber will store the image data from the camera in on-board, or system memory, sampling and digitizing analog data as necessary. In some cases the camera may output digital data, which is compatible with a standard computer interface like USB 2.0 or IEE-1394 Fire Wire, so a separate frame grabber may not be needed. The computer is often a PC or Macintosh and should be as fast as possible to keep the time needed to process each image as short as possible, or to allow more processing to be done in the time available. Machine vision software is needed to create the program which processes the image data. This may come in many forms, including C libraries of device drivers and functions, ActiveX controls, and point and click programming environments which allow easy assembly of image processing operations. When an image has been analyzed the system must be able to communicate the result to control the process or to pass information to a database. This requires a digital I/O interface or network connection. The human eye and brain can identify objects and interpret scenes under a wide variety of conditions. Machine vision systems are far less versatile so the creation of a successful system requires careful consideration of all elements of the system and precise identification of the goals to be accomplished, which should be kept as simple as possible.

References

1. Wilson, J. and Hawkes, J. F. B., *Optoelectronics: An Introduction,* Prentice-Hall International, London, 1983.
2. Jenkins, F. A. and White, H. E., *Fundamentals of Physical Optics,* 4th ed., McGraw-Hill, New York, 1981.

3. Hewitt, H. and Vause, A. S., *Lamps and Lighting,* Edward Arnold, London, 1966.
4. Fraden, J., Pyroelectric thermometers, in *The Measurement, Instrumentation and Sensors Handbook,* Webster, J. G., Ed., CRC Press, 1999, chap. 32.
5. Schuermeyer, F. and Pickenpaugh, T., Photoconductive sensors, in *The Measurement, Instrumentation and Sensors Handbook,* Webster, J. G., Ed., CRC Press, 1999, chap. 56
6. Sze, S. M., *Semiconductor Devices,* John Wiley and Sons, New York, 1985.
7. Ray, S. F., *Applied Photographic Optics,* 2nd. ed., Focal Press, Oxford, 1994
8. CCIR, *Characteristics of Monochrome and Colour Television Systems,* Recommendations and Reports of the CCIR, Vol. XI, Part 1: Broadcasting Service (Television), Section IIA, 1982.
9. Amelio, G. F., Charge coupled devices, *Scientific American,* 176, 1974.
10. Sheu, L. and Kadekodi, N., Linear CCDs, Advances in linear solid-state sensors, *Electronic Imaging,* August, 72, 1984.
11. Rutherford, D. A., A new generation of cameras tackles tomorrow's challenges, *Photonics Spectra,* September, 119, 1989.
12. Asano, A., MOS sensors continue to improve their image, *Advanced Imaging,* 42-44f, 1989.
13. Dierickx, B., Meynants, G., and Scheffer, D., Near 100% fill factor in CMOS active pixels, in *Proc. IEEE Workshop on Charge-Coupled & Advanced Image Sensors,* Brugge, Belgium, P1, 1997.
14. Ricquier, N. and Dierickx, B., Pixel structure with logarithmic response for intelligent and flexible imager architectures, *Microelectronics Engineering,* 19, 631, 1992.

19.9 Integrated Microsensors

Chang Liu

Introduction

The purpose of this section is to provide the general audience in the mechatronics field with information about micro-integrated sensors. It is my wish that an avid reader interested in the sensors area would be able to understand common fabrication techniques and sensing principles, and develop rudimentary background to guide the selection of commercialized sensors and development of custom sensors in the future.

Contents for this section are organized as follows. First, the general history of microsensors is discussed. This is followed by a brief discussion about major fabrication methods for microsensors. Commonly used principles for sensors are reviewed next. Sensing of a physical parameter of interest can be achieved using various structures and under different sensing principles. Examples of sensors, along with their structures and fabrication techniques, are provided to familiarize the readers with the configurations and fabrication methods for each.

The microsensors research area covers diverse disciplines such as materials, microfabrication, electronics, and mechanics. A comprehensive coverage of all aspects is beyond the scope of this book. We will focus on a few primary sensing principles and frequently used sensors examples. A reader would be able to grasp a glimpse of the sensors field from the perspectives of sensing principles and of application areas. References for further in-depth studies are provided when appropriate.

Definition of Integrated Microsensors

Integrated microsensors refer to sensors or arrays of sensors that are developed using microfabrication technology. The characteristic length scale of individual microsensors ranges between 1 μm and 1 mm. Nanosensors refer to sensors with characteristic length scale on the order of 1 nm to 1 mm. In this text, we are mainly concerned about sensors for detecting physical variables such as force, pressure, tactile contact, acceleration, rotation, temperature, and acoustic waves. Chemical sensors, used for sensing the concentration of chemicals or pH values, are beyond the scope of this book.

A brief historical overview of microsensors development is presented here. The microsensors are made possible by using integrated microfabrication technology, first developed for making integrated circuits. Since the invention of the first transistor in 1947 and the successful demonstration of the integrated circuits in 1971, technologies and equipment for building integrated and miniature circuit components on semiconductor substrates (e.g., silicon and GaAs) have improved rapidly. The integration density doubles every 12–18 months following the empirical Moore's law. The integrated circuit technology revolutionized the modern society by enabling low cost analog signal processors, digital logical units, computer memories, and CCD cameras. These achievements should serve as evidence of the power of integration and miniaturization technology.

Advanced signal processors such as analog ASIC (application specific integrated circuits) and CPU (central processing units) are merely one aspect of a highly intelligent mechtronics system. Sensors are of critical importance for mechantronics systems to interact with the physical world. In the 1970s, a few researchers experimented with using IC fabrication technology to realize mechanical transduction elements on a silicon chip. H. Nathansan [1] developed floating gate transistors where the gate is made of a suspended cantilever beam and its distance to the conducting channel can be adjusted using electrostatic forces. Work by several pioneering researchers resulted in the first commercial pressure sensors [2], accelerometers [3], integrated gas chromatometers, as well as the ink jet printer nozzle array [4,5].

There are several important advantages associated with integrated microsensors compared with conventional macroscopic sensors. Miniaturization of sensors means that such sensors offer better spatial resolution. In many cases, reduced inertia and thermal mass translate into higher mechanical resonant frequency and lower thermal time constants. Since such sensors are fabricated using photolithography methods, their costs can be low, as many identical units are made in parallel (if the demand is sufficiently large). Further, since the geometric features of sensor components are defined by precision photolithography, the uniformity and repeatability of the performance of such sensors are significantly improved over conventional sensors. The capability to monolithically integrate sensors and integrated circuits reduces the path length between sensors and circuits and increases the signal-to-noise ratio.

Fabrication Process of Integrated Microsensors

As mentioned previously, the microfabrication technology for microsensors was developed based on integrated circuits-compatible platforms. As a result, silicon has historically been a predominant material for microintegrated sensors. In other words, a majority of sensors are now made with silicon wafer as a substrate. However, in recent years, new materials such as polymers are being applied to microfabrication. Polymer materials offer lower costs than single crystal silicon wafers and, in some cases, simpler processability, compared with semiconducting silicon. There are few examples of sensors that are entirely based on polymer these days, because certain key sensing elements are not yet available in polymer format. However, with the advancement of polymer microfabrication technology and organic semiconductors, all-polymer sensors can be predicted for future use.

Microsensors and mechanical elements (notably cantilevers and diaphragms) can be made from silicon substrates in a variety of ways. The two primary categories of fabrication methods are called bulk micromachining and surface micromachining. In bulk machining, a portion of the silicon substrate is removed using chemical wet etch or plasma-assisted dry etch to render freestanding mechanical members. For silicon substrates, the following wet chemical etchants are frequently used: potassium hydroxide (KOH), ethylene-diamine pyrocatecol (EDP), or tetramethyl ammonium hydroxide (TMAH). Dry etching methods use AC-excited plasma as an energetic source to selectively and anisotropically remove the substrate materials. A review of etching solutions commonly used in the silicon microfabrication industry and their respective etch rates on various materials can be found in [6].

In surface micromachining methods, a freestanding structure typically resides within a thin region near the substrate surface. The fabrication process involves only layers on the surface of a substrate, hence the name surface micromachining. The fabrication process is typically referred to as a sacrificial etching method as well. First, a thin-film solid layer called the sacrificial layer is placed on the wafer surface. This is followed by the deposition of a structural layer, which constitutes the mechanical structure (e.g.,

cantilever or membrane). An etchant that selectively etches the sacrificial layer with a much greater rate than the structural layer is used to remove the sacrificial layer without damaging the structure layer, leaving the structural material freestanding.

In the ensuing section, the fabrication process related to specific examples of sensors will be discussed to illustrate specific uses of bulk and surface micromachining methods. Both surface and bulk micromachining techniques offer advantages and disadvantages. For example, the surface micromachining process is generally compatible with established integrated circuit foundries because no substrate etching is involved. The bulk micromachining also involves somewhat lengthy substrate removal. However, a bulk silicon micromachining process is capable of realizing single crystal silicon structures with extremely low intrinsic mechanical stress and bending. For development of custom sensors, the selection of a fabrication process must be done carefully to achieve desired device characteristics and fabrication yield and to reduce overall sensor costs.

Resources Regarding Sensors

The community that develops microfabrication methods and microintegrated sensors frequently publishes archival results in the *Journal of Microelectromechanical Systems (MEMS), Journal of Sensors and Actuators*, and the *Journal of Micromechanics and Microengineering*. Major conferences in this area include: (1) the IEEE workshop on solid-state sensors and actuators (held biannually at the Hilton Head island, South Carolina); (2) the International Conference on Solid-State Sensors and Actuators (held biannually at international venues); and (3) International Conference on Micro-Electro-Mechanical Systems (held annually at international venues).

Interested readers can find more in-depth discussions on microsensors in a number of reference books [7,8] and websites [9].

Examples of Micro- and Nanosensors

Basic Sensing Principles

Microsensors are based on a number of transduction principles, including electrostatic, piezoresistive, piezoelectric, and electromagnetic (including optical sensing). The fundamental principles of these sensing methods are discussed in the following.

Electrostatic Sensing

In electrostatic sensing, a physical variable of interest, such as force or vibration, is transduced into mechanical displacement of a cantilever beam or a membrane. A schematic diagram of a typical transduction structure is shown below (Fig. 19.112). The moving cantilever or membrane forms a capacitor with a reference, typically immobile, electrode. For the structure shown in the diagram below, the displacement of the top plate induces changes of the capacitance. The electrostatic sensing is also commonly referred to as capacitive sensing. The changes in capacitance are used to provide information about the parameter of interests. The electrostatic sensing principle can be used for accelerometers, acoustic sensors, rotation gyros, pressure sensors, tactile sensors, and infrared sensors. Respectively in these examples, the external excitations responsible for member movements are inertia force, air mass vibration, Coriolis force, pressure, contact force, and thermal bimetallic bending due to absorbed energy and increased temperature.

The electromechanical model of a simple capacitive sensor shown in Fig. 19.112(a) is illustrated in Fig. 19.112(b). The top electrode is supported by two suspension beams with a combined equivalent force constant (spring constant) of k. The capacitance value of a parallel plate capacitor is expressed as

$$C = \frac{\varepsilon_r \varepsilon_0 A}{d}$$

where A is the area of the electrode, d is the distance between two electrodes, ε_r and ε_0 are the relative permitivity of the media and the permitivity of vacuum, respectively. When the distance between the

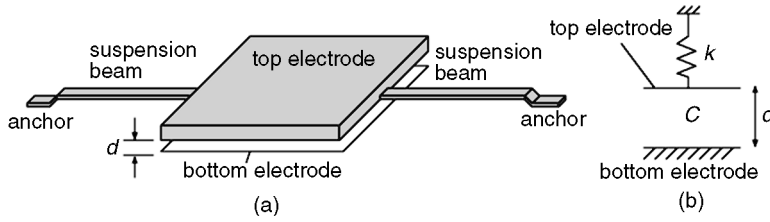



FIGURE 19.112 Schematic diagram of a parallel plate capacitor: (a) perspective view, (b) electromechanical model.

two electrodes changes by Δd, the first-order estimate of the change in capacitance is

$$\Delta C = \frac{C}{d}\Delta d$$

When a force F is applied to the top plate, the change of capacitance value is

$$\Delta C = \frac{C}{d}\frac{F}{k}$$

Piezoresistive Sensing

In a piezoresistive sensor, the magnitude of a mechanical displacement is measured by the amount of stress it induces in a mechanical member. A stress-sensitive resistor (called a piezoresistor) located strategically on the mechanical member experiences a change of resistance as a result of the applied stress. Many materials, including metals, alloys, and doped silicon, exhibit piezoresistive characteristics. The applied stress causes the lattice of a material to deform, thereby inducing changes in the resistivity as well as the dimensions of a resistor. The change of resistance (ΔR) as a function of applied strain ε is

$$\frac{\Delta R}{R_0} = G\varepsilon$$

where R_0 is the value of the resistor in the unstressed state, and G is the piezoresistive gauge factor.

Using doped silicon as a piezoresistive sensor, the overall footprint of the sensor can be made quite small and yet have a respectable value, i.e., 1 kΩ. Unlike the capacitive sensor method, which requires significant plate area to achieve significant capacitance value, the piezoresistive sensor is more area efficient. As a result, the piezoresistive sensing is more likely to be used for sensors with characteristic length below 10 μm. However, the capacitive measurement method is more generally applicable, whereas the optimal piezoresistive sensors involve silicon with proper doping concentration.

Piezoelectric Sensing

A piezoelectric material is one that produces electrical polarization (internal electric field) when an external mechanical strain is applied. A piezoelectric material also exhibits reverse piezoresistivity. Namely, a mechanical strain (or displacement) will result when a voltage (or electric field) is applied on the material itself. The reverse piezoelectricity is commonly used as an actuation principle for producing mechanical movement or force.

One advantage of piezoelectric sensing over piezoresistive sensing lies in the fact that piezoelectric sensors are self-generating, i.e., a potential difference will be created without external power supply. However, high quality piezoelectric films with consistent and uniform performance characteristics require dedicated machinery and calibrated processes. Such a technical barrier prevents thin film piezoelectric materials to be as widely used as piezoresistive elements. However, high quality piezoelectric films are

increasingly becoming available through commercial services. Commonly used piezoelectric materials in microfabricated sensors include sputtered zinc oxide and lead zirconate titanate (PZT).

Temperature Sensing

Temperature sensing is used not only for measuring the temperature of the ambient but also for inferring heat transfer. Temperature sensors can be made of thermal resistors or thermal couples. A thermal resistor is a thin film resistive element whose resistance changes as a function of the temperature. This is explained by changes in resistivity as well as dimensions. Doped semiconductor materials (e.g., single crystal silicon or polycrystalline silicon) exhibit temperature coefficients of resistors (TCR, denoted α) on the order of –0.1%/°C to 5%/°C. For a thermal resistor, the normalized changes in resistance (R) are related to the change in temperature (T) by

$$\frac{\Delta R}{R} = \alpha T$$

Thermal couples are made of two different materials with different Seebeck coefficients. The voltage induced by a single thermal couple junction is proportional to the difference in temperatures at the junction and of the ambient. For more information about thermal couples, readers can refer to references.

Temperature sensing is also commonly achieved using a thermal bimetallic beam. A composite beam made of two materials with different thermal expansion coefficients will bend as the two parts expand with different speed. The amount of mechanical bending, sensed electrostatically or using piezoresistors, corresponds to the applied temperature. Such sensing principle has been used for making uncooled infrared sensors.

Pressure Sensors

Pressure sensors are important for industrial and automotive control and monitoring. Existing micropressure sensors consist of diaphragms that deform in response to pressure differences. Using micromachining technology, the diaphragms can be made very thin, hence greatly increasing the sensitivity of sensors over conventional pressure sensors with thick diaphragms. Integrated microfabrication technology also enables sensors to be made in conjunction with signal-processing circuits. Since the distance between the sensor diaphragm and the signal processors are close, the noise is generally much lower compared with conventional sensors.

The diaphragm is a critical element in a pressure sensor. It can be made by either surface micromachining or bulk micromachining techniques. Hence, we classify micropressure sensors according to the methods of forming the diaphragms. In each category of pressure sensors, the displacement of the diaphragm can be determined by using piezoresistive sensing or capacitive sensing.

Over the past two decades, many micromachined pressure sensors have been developed, some commercialized successfully for automotive and machinery applications. The intent of this section is not to present an exhaustive summary of all the work that has been accomplished, but rather discuss several representative devices with the purpose of (1) providing the readers with a general overview of the available technologies; and (2) providing leads to the existing body of literature in this area.

Bulk Micromachined Pressure Sensors

The schematic diagram of a bulk micromachined pressure sensor is illustrated in the diagram below. The diaphragm will bend when a differential pressure is applied across it. A common technique for sensing the diaphragm displacement is by using piezoresistors embedded in the diaphragms. However, it should be noted that other sensing principles are also feasible.

In the diagram below (Fig. 19.113), four piezoresistors are embedded near the edge of the diaphragm. The locations of sensors are selected carefully such that under a given displacement at the center of the diaphragm, the magnitude of the stress is the greatest at the sensor locations. There are a number of choices for the diaphragms and the piezoresistors. A number of possible materials and their relative merits are summarized in the Table 19.7. Three distinct pressure sensor architectures and their respective fabrication processes are discussed in the following paragraphs.

TABLE 19.7 A List of Possible Materials for Diaphragm and the Piezoresistive Sensors

Diaphragm material	Piezoresistor material	Relative merits
Single crystal silicon	Doped single crystal silicon	Relatively difficult to control the thickness of the diaphragm
Silicon nitride thin film	Polycrystalline silicon	Easy to form thin diaphragms; involved LPCVD polysilicon

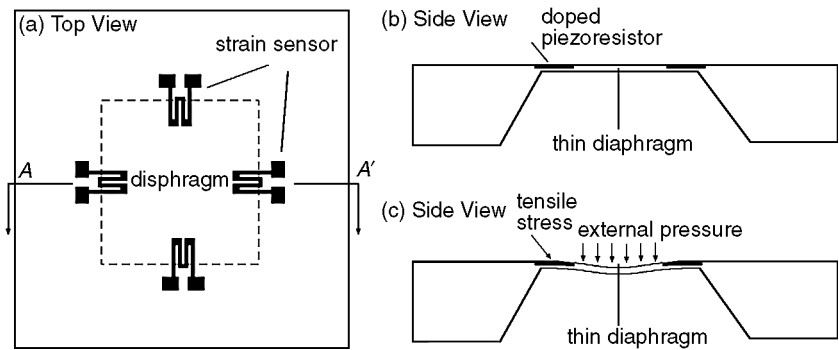



FIGURE 19.113 Schematic diagram of a bulk micromachined pressure sensor.

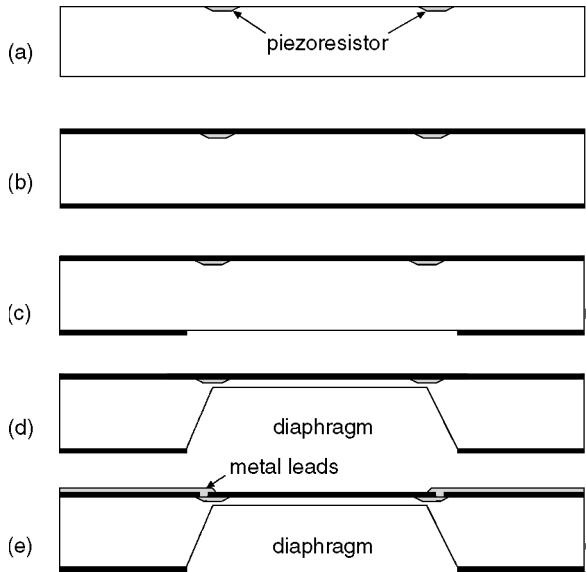



FIGURE 19.114 Schematic diagram illustrating major steps in the microfabrication process of a bulk micromachined pressure sensor.

The fabrication process for a pressure sensor using plain silicon wafer as the substrate is shown in Fig. 19.114. In the first step, the wafer is selectively doped with boron or phosphorous atoms to create piezoresistors on the front side (a). The wafer is then passivated with a thermally grown silicon dioxide thin film (b). In the ensuing step, the silicon dioxide film on the backside is patterned and selectively etched to expose the silicon (c). The exposed silicon material will be etched when the wafer is immersed in an anisotropic silicon etchant (d). In order to form the silicon diaphragm with desired thickness,

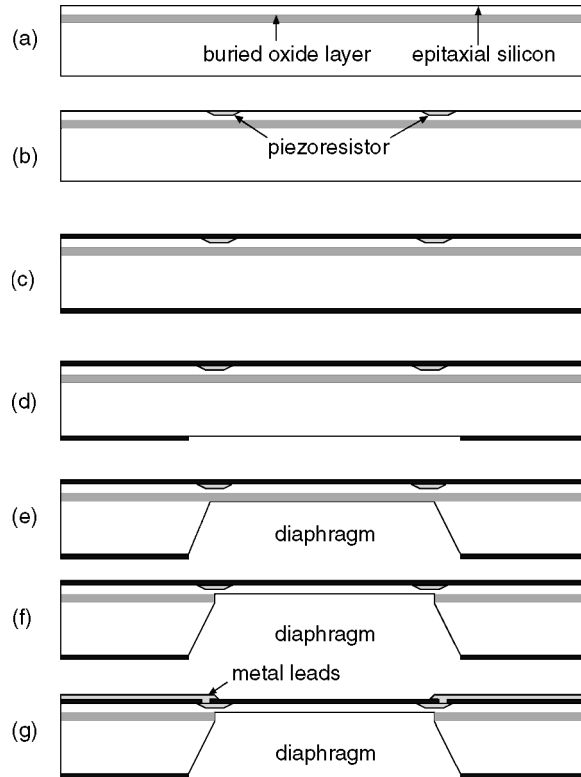



FIGURE 19.115 Schematic diagram of an alternative process for realizing bulk micromachined pressure sensors.

workers resort to timed etch with precise knowledge of calibrated wafer thickness and etch rates. However, this step must be performed with caution as the etch rate may vary with time and locations on a wafer. Typically, the resultant thickness of the diaphragm is large (30–50 μm) to ensure sufficient yield of devices. In the final step, the oxide on the front side of the wafer is patterned to provide contact vias for metal lead wires (e).

To circumvent the problem of process uncertainty of the aforementioned process, wafers with barrier layers can be used (Fig. 19.115). For example, it is possible to use a silicon-on-insulator (SOI) wafer with a thin film of silicon on top of a silicon dioxide layer. The silicon and the oxide layers lie on top of the bulk silicon substrate (a). Following steps similar to the ones discussed above, one can form piezoresistors (b) and open windows in silicon oxide on the backside of the wafer (d). The anisotropic etchant of silicon has minimal etch rate on the silicon oxide, hence the through-wafer etch will automatically stop when the buried oxide layer is exposed. This allows a professional engineer to perform adequate overetch to ensure that diaphragms on all devices reach the same thickness (e). This self-limiting etching behavior reduces the complexity of process control and is conducive to reducing the process costs. The oxide layer is then selectively removed using hydrofluoric acid, which does not etch silicon. Hence a thin silicon diaphragm, with the thickness defined by the thickness of the epitaxial silicon layer specified during the SOI wafer manufacturing, can be formed efficiently. Finally, via holes are opened on the frontside and metal leads are deposited and patterned (g).

Although this process is advantageous over the one introduced earlier, it has a few shortcomings. For example, although the process discussed above is much more efficient in terms of controlling the diaphragm thickness, the SOI wafers used in the process are more expensive than ordinary silicon wafers. Even with SOI wafers, the thickness of the silicon diaphragm is typically 2–10 μm. In order to further

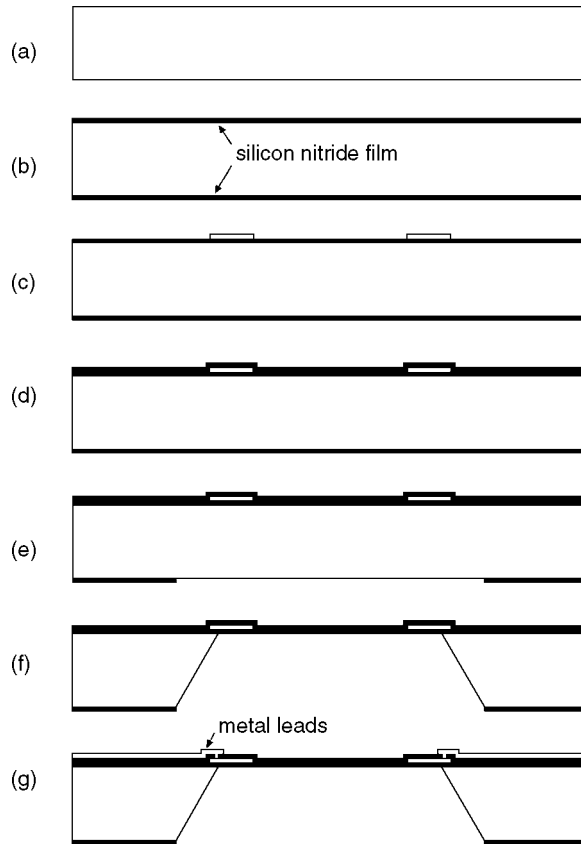



FIGURE 19.116 Schematic diagram of major process steps for realizing a bulk micromachined pressure sensor with a silicon nitride diaphragm.

increase the pressure sensitivity, it is advantageous to reduce the thickness of the diaphragm further. However, this would be difficult to achieve if silicon is the diaphragm material. In the following, a process of using silicon nitride thin film will be discussed.

An alternative sensor structure uses silicon nitride thin film as the diaphragm and deposited and doped polycrystalline silicon as the piezoresistive sensor. The process is described in Fig. 19.116. Starting with a bare silicon wafer (a), a layer of silicon nitride film is deposited using LPCVD methods (b). The wafer is then coated with a layer of polysilicon with suitable doping concentration (c). The polysilicon is patterned and defined. This is followed by the deposition of yet another thin film silicon nitride to protect the polysilicon film during the ensuing silicon etching (d). The thickness of the two LPCVD silicon nitride layers is the thickness of the finished diaphragm. A window is opened on the backside of the wafer to expose the silicon material. The silicon is etched in an anisotropic etchant, which does not attack the silicon nitride film. In other words, the selectivity between silicon and silicon nitride is high. Following the formation of the diaphragm, the silicon nitride on top of the polysilicon resistors is selectively patterned and metal leads are formed (g).

Surface Micromachined Pressure Sensors

The surface micromachining process does not require the removal of silicon substrate, which is time-consuming and not fully compatible with integrated circuit processes at the present because of the silicon etchants used. For producing low-cost, high-performance integrated sensors, surface micromachining

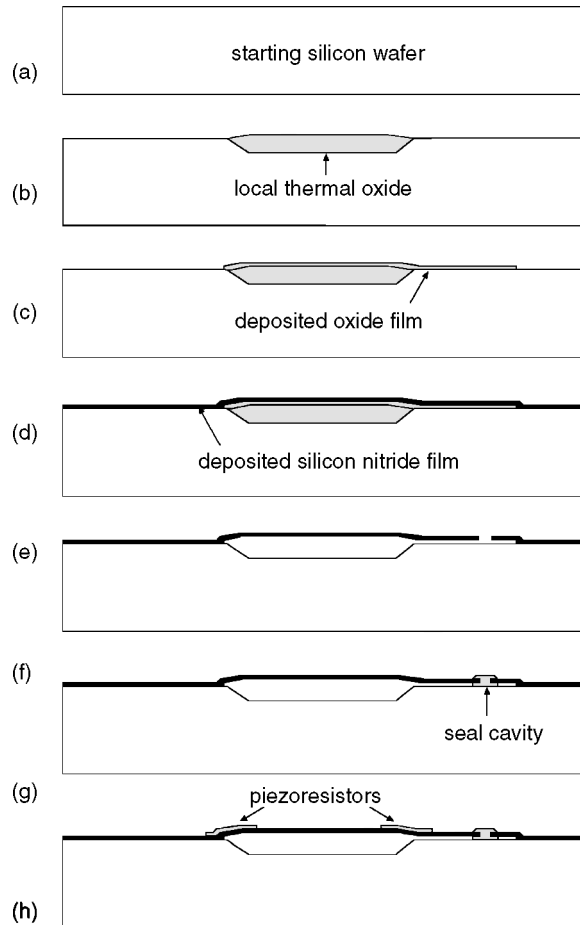



FIGURE 19.117 Schematic diagram of major steps for making a surface micromachined pressure sensor with silicon nitride diaphragm.

offers important advantages. An examplary surface micromachining process is described in the following paragraph.

The fabrication process for a surface micromachined pressure sensor is shown in Fig. 19.117. It starts with a silicon substrate (a) with the front side polished. A local thermal oxidation process is performed first to form a silicon oxide well with a typical thickenss of 1.3 μm. The thermal oxide is part of the sacrificial layer that will be removed at a later stage. Using a process called low-pressure chemical vapor deposition (LPCVD), a thin layer of oxide is again deposited over the wafer surface. This oxide layer is patterned using the photolithography method (c). The entire wafer is coated with a silicon nitride thin film deposited by LPCVD technique as well (d). The silicon nitride film is patterned and etched to produce an access hole on top of the underlying oxide layer (e). Through this access hole, hydrofluoric acid removes the oxide materials inside the cavity. The etch rate of the acid on silicon nitride is negligible (f). After the cavity is emptied and dried, another layer of LPCVD silicon nitride is deposited to seal the opening in the original silicon nitride layer (g). Following this step, polycrystalline silicon with suitable doping concentration is deposited on top of the wafer and patterned to form the piezoresistors (h).

It should be noted that piezoresistive sensing, though dominant in the methods reviewed, is not the only sensing mechanism available. Capacitive sensing and piezoelectric sensing are also feasible and have been demonstrated in the past. However, discussions of these methods are beyond the scope of this text.

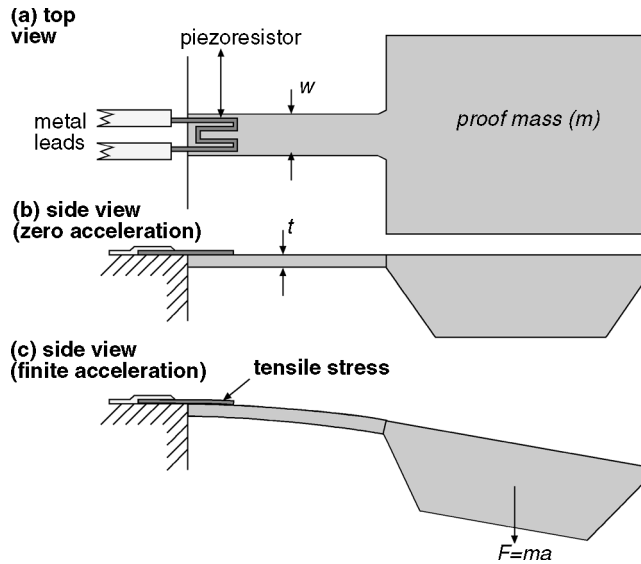



FIGURE 19.118 Schematic diagram of a bulk micromachined accelerometer.

Pressure Sensors Made of Non-Silicon Materials

For certain applications such as monitoring of internal combustion engine, pressure sensors are required to sustain high temperature of operation. In such cases, silicon is not the optimal material because high temperature causes doped silicon junctions to fail.

Work has also been done to implement polymer materials for pressure sensors. Though such devices are relatively few, they represent an important development trend for future sensors.

Accelerometers

Bulk Micromachined Accelerometers

Acceleration sensors (or so-called inertial measurement units, IMU) are important for monitoring acceleration and vibration experienced by a subject, such as an automobile, a machine, or a building. Low-cost accelerometers used in automobile airbag deployment systems can reduce the costs and enhance driver safety. Micromachined sensors can be made small and sufficiently low-cost to be used in smart projectiles, for example, concrete penetrating bombs. Small, multi-axial accelerometers can also be applied in writing instruments (smart pens) for handwriting recognition.

A representative bulk micromachined accelerometer is illustrated in Fig. 19.118. A SEM micrograph of a prototype sensor is shown in Fig. 19.119. A silicon proof mass is attached to the end of a cantilever beam. At the base of the cantilever beam lies a piezoresistive element. Supposing the mass of the proof mass is m, and the magnitude of the acceleration is a, one can estimate the sensor output following a few simple analysis steps. First, a concentrated force with a magnitude of $F = ma$ is applied in the center of the proof mass according to Newton's first law. Secondly, the force translates into a torque loading at the base of the cantilever with the magnitude being

$$M = F\left(l + \frac{L}{2}\right) = ma\left(l + \frac{L}{2}\right)$$

The magnitude of the strain experienced at the surface of the cantilever beam, where the piezoresistors are located, is

$$\varepsilon = \frac{Mt}{2EI}$$

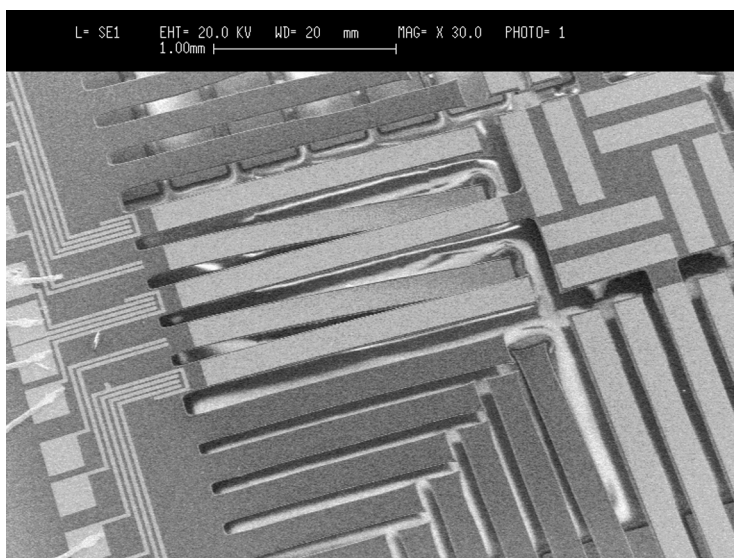



FIGURE 19.119 A SEM micrograph of a prototype bulk micromachined pressure sensor (Junjun Li).

Here, the term t is the thickness of the beam, E is the modulus of elasticity of the cantilever beam material, and I is the momentum of inertia associated with the beam cross section. Supposing the cross section of the cantilever beam is a rectangle with a width w and a thickness t, the moment of inertia is

$$I = \frac{wt^3}{12}$$

Note that the moment of inertia is strongly related to the thickness of the beam. If the thickness of beam is reduced to half, the magnitude of I is reduced by eight times, and the sensitivity of the sensor increases by eight fold.

Surface Micromachined Accelerometers

Surface micromachined accelerometers offer the potential advantage of ready integration with signal processing circuits. As a result, various types of surface micromachined versions have been made in the past decade. A successful commercial product has been made by analog devices for sensing automobile acceleration to deploy airbags in the events of collision. The structure, operational principle, and fabrication process for such a sensor is briefly discussed in this section.

The sensor consists of two sets of interdigited comb-finger-shaped electrodes as shown in Fig. 19.120. One set of fingers is stationary and fixed to the substrate. Another set if suspended by cantilever springs to the substrate. Capacitors are formed between each pair of comb-like fingers. When an external acceleration is applied along the horizontal axis, an inertia force is applied to the moving set of fingers and causes the moving fingers to displace. The amount of displacement is related to the magnitude of the acceleration and the force constant of the supporting springs. The relative motion of the two sets of fingers result in changes of the overall capacitance value between the two sets of fingers. The minute capacitance change is sensed and processed by a signal-processing circuit consisting of an $\Sigma - \Delta$ A/D conversion stage [10].

The fabrication process for such a sensor according to the A-A cross-section is illustrated in Fig. 19.121. First, transistors for signal processing circuits are first made on a silicon substrate (a). A sacrificial silicon dioxide layer is deposited onto the wafer surface (b), followed by the deposition of a polycrystalline silicon

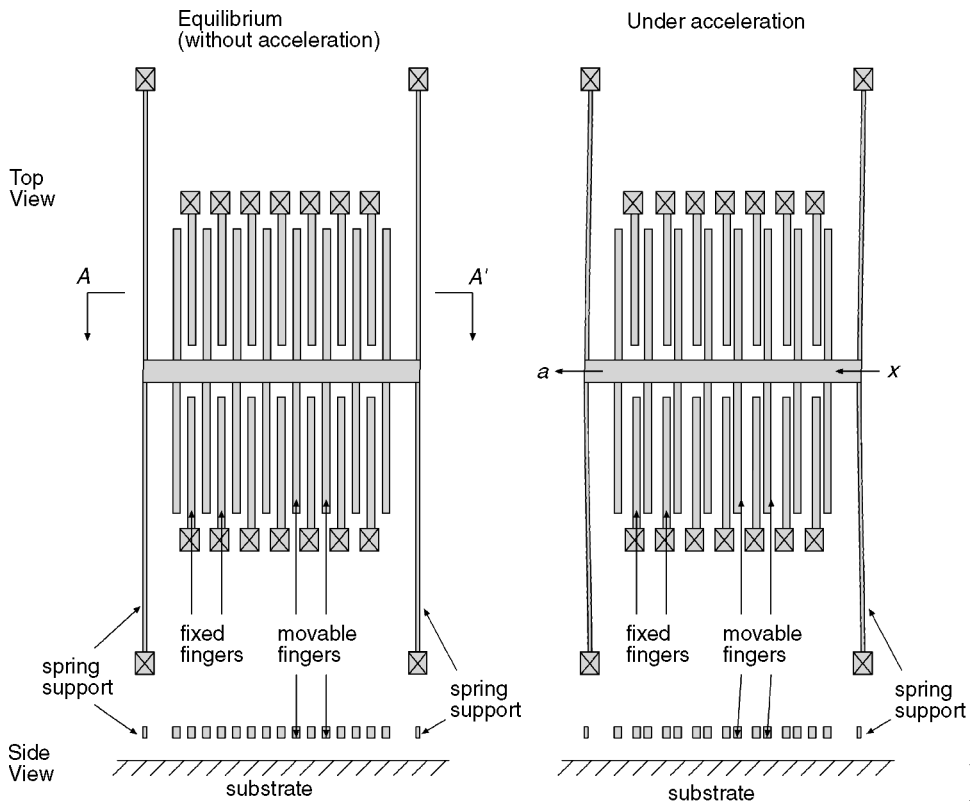



FIGURE 19.120 Schematic diagram illustrating the operation principle of a surface micromachined accelerometer.

layer (c). The polycrystalline silicon is patterned and etched to form the comb fingers (e). Subsequently, the oxide layer is removed by using a web etchant (hydrofluoric acid) that etches polycrystalline with negligible rates. In areas where the polysilicon is anchored to the substrate, a via hole is patterned and etched in the sacrificial layer before step (c).

Tactile Sensors

Tactile sensors are most widely used for robotics applications to provide tactile sensations for object handling. The sensor density on a human fingertip is on the order of 100/cm^2. Such a high sensor density can be achieved using microfabrication technology.

An arrayed tactile sensor is illustrated in Fig. 19.122. A two-dimensional array of individual sensor elements provides two-dimensional mapping of contact force and shear force. The schematic cross-sectional diagram of an array in contact with an arbitrary object is shown in Fig. 19.122(b). As an object contacts a sensor beam, the amount of displacement corresponds to the contact force as well as the surface topology.

The fabrication process of the tactile sensor is discussed in the following and illustrated in Fig. 19.123. Starting with a silicon wafer (a), a local ion implantation is first conducted to produce piezoresistors (b). A thermal oxide film is grown to provide passivation to the entire wafer. The oxide layer on the bottom of the wafer is patterned and etched to expose silicon substrates (c). An anisotropic silicon etch is performed to remove silicon from the backside of the wafer (d). The oxide film on the front of the wafer is then patterned and etched using plasma anisotropic etch to create free-standing cantilever beams (e and f). Metal thin film is then deposited and patterned to provide lead wires (g).

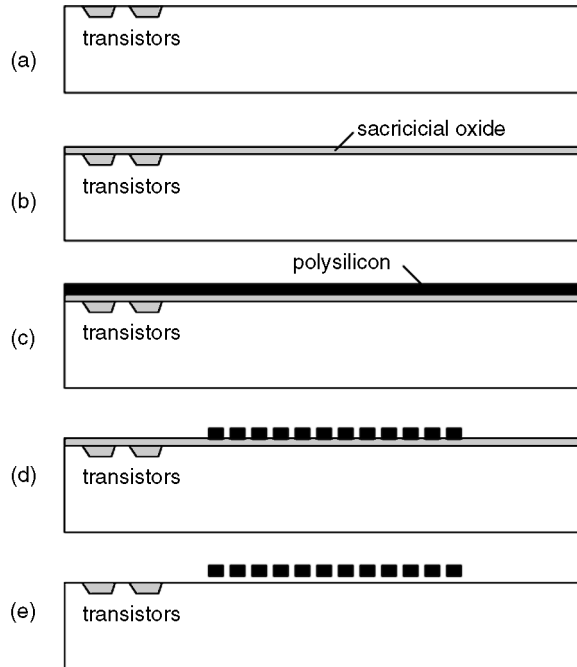



FIGURE 19.121 Schematic diagram of the fabrication process for a surface micromachined accelerometer illustrated in the previous figure.

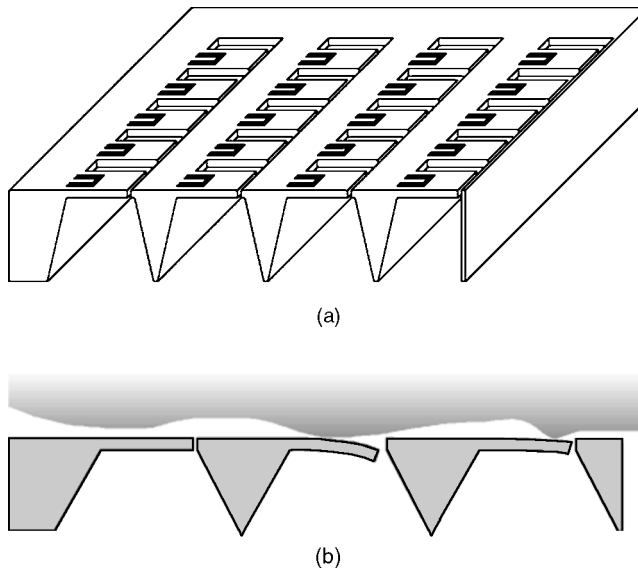



FIGURE 19.122 Schematic diagram of an array tactile sensor: (a) perspective view, (b) cross-sectional view.

Flow Sensors

Sensors for monitoring the flow rate of fluid (air or liquid) and for measuring the drag force exerted on an object moving in a fluid have important applications in robotics applications. Existing flow sensors are based on a number of principles, notably thermal and momentum transfer principles.

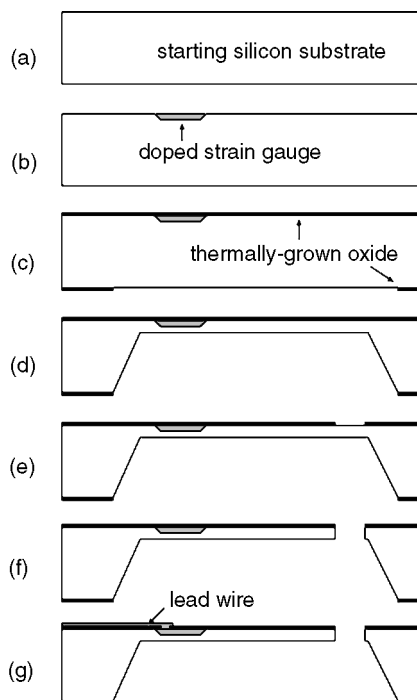



FIGURE 19.123 Schematic diagram of the microfabrication process for realizing a tactile sensor.

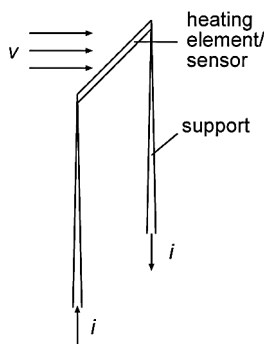



FIGURE 19.124 Perspective diagram of a thermal-transfer based flow sensor (anemometer).

Flow Sensors Based on Heat Transfer Principles

For sensors based on thermal transfer principles, a heated element is used with temperature slightly above the temperature of the ambient fluid (Fig. 19.124). The heat is generally created by passing current through a resistive element. An ideal element to serve as the heating element is doped polysilicon resistor. The resistivity is generally lower than what can be achieved using metal resistors of the same dimension, hence the resistance value is greater and the heating element can be made smaller.

The movement of the fluid creates velocity-dependant forced convection of heat, thereby reducing the temperature of the heated element accordingly. The temperature of the element is therefore used to provide information about the flow rate and direction. Such sensors are commonly referred to as hot-wire anemometers. Micromachined hot-wire anemometers have been demonstrated by several groups [11,12].

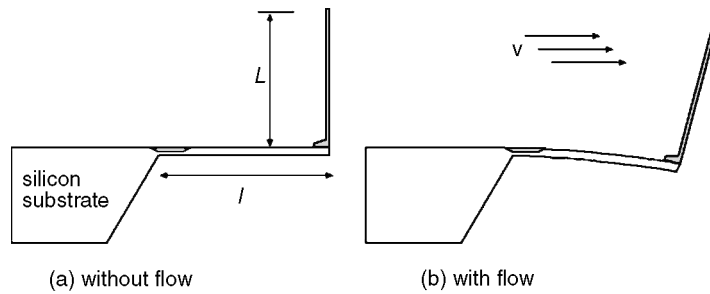



FIGURE 19.125 Schematic diagram illustrating the operation principle of a momentum-transfer based flow sensor.

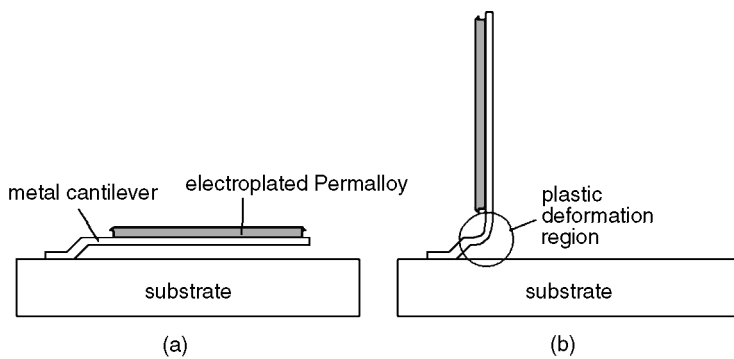



FIGURE 19.126 Schematic diagram of the plastic deformation magnetic assembly (PDMA) process.

Flow Sensors Based on Momentum Transfer Principles

For sensors based on momentum transfer principles, a mechanical member is bent by the momentum imparted by a moving fluid (Fig. 19.125). The amount of the bending is used to decipher the strength of the fluid flow. The schematic diagram of an exemplary flow sensor is shown in the figure below. It consists of a vertical shaft attached to the end of a cantilever beam (a). When an external flow is exerted, it will apply a distributed force onto the vertical shaft, hence causing the cantilever to bend. The extent of the bending, as sensed by the embedded piezoresistor, is proportional to the average flow rate.

The fabrication process is similar to the tactile sensor except for the attachment of the integrated vertical shaft. A number of techniques for assembling three-dimensional microstructures using efficient integrated processes have been developed in the past. For example, three-dimensional structures can be realized using hinged microstructures and using solder joints or polymer joints. Recently, a process called the plastic deformation magnetic assembly (PDMA) has been developed. In the following paragraph the PDMA process is briefly discussed.

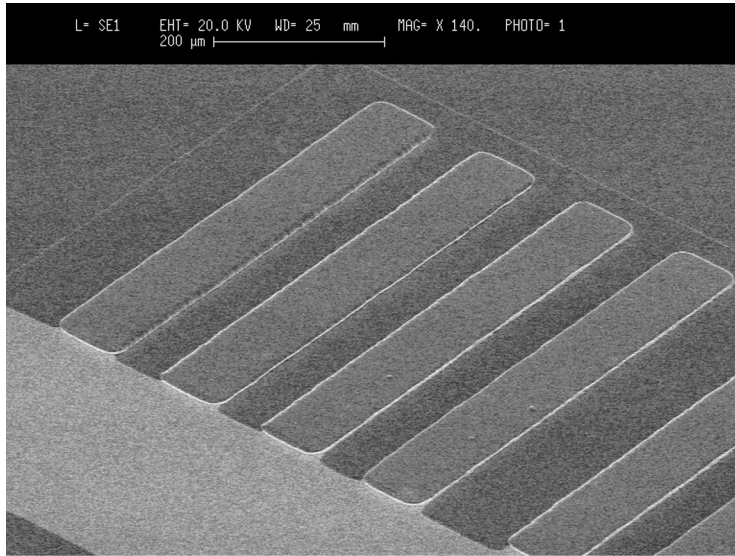
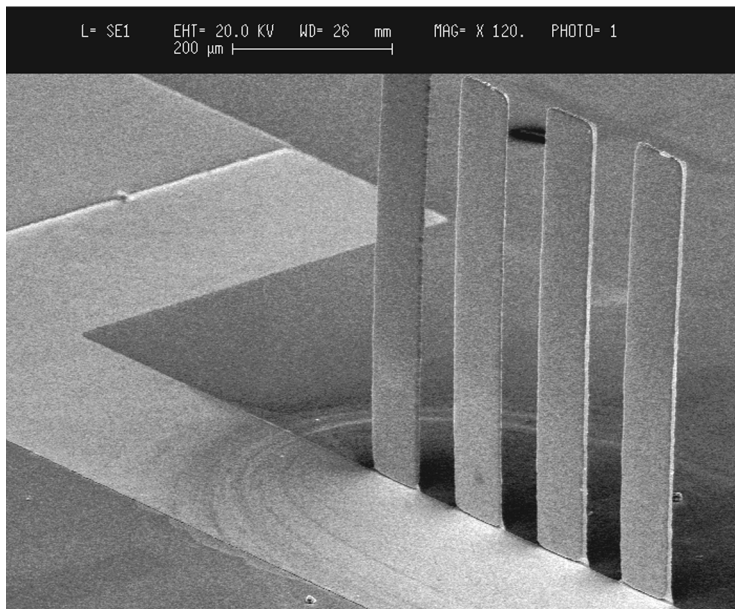
The PDMA technique is discussed using a simple surface micromachined cantilever as an example. As shown in the diagram below (Figs. 19.126 and 19.127), a single-clamped cantilever made of a ductile metal (e.g., gold or aluminum) is suspended from the substrate. A piece of Permalloy, a ferromagnetic alloy made by electrodeposition, is attached to the cantilever. When a magnetic field is applied from underneath the wafer, the magnetic piece will be magnetized and will experience a magnetic torque M. The torque lifts the cantilever beam away from the substrate. If the amount of bending is significant, the ductile metal will be displaced permanently due to plastic deformation at the hinge region.

(a)

(b)

FIGURE 19.127 SEM micrographs of cantilever beams (a) while in plane, and (b) after PDMA assembly.

Future Development Trends

Miniaturization and integration of circuits has resulted in revolution in the society so far. It drastically reduces the costs and increased the performance of circuits. Without the integrated circuits technology, the information age would not have dawned on the human society.

It is conjectured that integrated microsensors are likely to produce as broad and deep an impact on the society as the integrated circuits. Sensors can be used for robotics sensing, smart buildings, smart

toys, automotive safety and control, and industrial control. However, in order to realize the advantages of integrated sensors, a number of technical barriers must be overcome. Two important barriers are (1) high R&D costs of integrated sensors and (2) reliable and robust packaging of sensors.

The development of microintegrated sensors involves high development costs and long time-to-market. The surface micromachined accelerometer developed by analog devices costs tens of millions of dollars and took more than 5 years to produce. Why do integrated sensors cost so much to build? Sensors are developed using a group-up approach. The development cycle of a sensor begins at the level of physical principles. The cost of sensor development includes expertise for material selection, design generation, prototype process development, and characterization.

Such a development cost and speed is not tolerable in applications where only a small amount of custom sensors is required. Standard sensing modules, low-cost, flexible foundry fabrication processes, and advanced computer simulation and prototyping tools are required to advance the state-of-the-art of microintegrated sensors.

Future sensors will involve more non-silicon materials. For example, polymer materials can be used to reduce the costs while high-temperature materials maybe used for high-temperature sensing applications (e.g., monitoring of conditions in engines).

Conclusions

A brief historical overview of the development of microfabrication technology and microintegrated sensors is presented. Common sensing principles, including capacitive, piezoresistive, and piezoelectric sensing, are discussed. Four important case studies of sensors are undertaken. For each type of sensor applications—pressure sensors, acceleration sensors, tactile sensors, and flow sensors—the sensor architectures and fabrication processes are reviewed. Interested readers may find more in-depth information in the references provided in this section.

References

1. Nathanson, H.C., Newell, W.E., Wickstrom, R.A., and Davis J.R. Jr., “The resonant gate transistor,” *IEEE Transactions on Electron Devices,* Vol. ED-14, No. 3, pp. 117–113, March 1967.
2. Petersen, K.E., “Silicon as a mechanical material,” *Proceedings of the IEEE,* Vol. 70, No. 5, pp. 420–457, May 1982.
3. Angell, J.B., Terry, S.C., and Barth, P.W., “Silicon micromechanical devices,” *Scientific American,* Vol., 248, pp. 44–55, April 1983.
4. Siewell, G.L., Boucher, W.R., and McClelland, P.H., “The ThinkJet orifice plate: a part with many functions,” *Hewlett-Packard Journal,* May 1985, pp. 33–37.
5. Allen, R.R., Meyer, J.D., and Knight, W.R., “Thermodynamics and hydrodynamics of thermal ink jets,” *Hewlett-Packard Journal,* May 1985, pp. 21–27.
6. Williams, K.R., and Muller, R.S., “Etch rates for micromachining processing,” *Journal of Microelectromechanical Systems,* Vol. 5, No. 4, pp. 256–268, December 1996.
7. Kovacs, G.T.A., Micromachined transducers sourcebook, McGraw-Hill, 1998.
8. Trimmer, W.S., Micromechanics and MEMS—Classis and seminal papers to 1990, IEEE Press, 1997.
9. WWW site http://mems.isi.edu.
10 Yun, W., Howe, R.T., and Gray, P.R., “Surface micromachined, digitally force-balanced accelerometer with integrated CMOS detection circuitry,” *Technical Digest, IEEE Solid-State Sensor and Actuator Workshop,* pp. 21–25, Hilton Head, SC, June 1992.
11. Jiang, F., Tai, Y.C., Ho, C.M., Karan, R., and Garstenauer, M., “Theoretical and experimental studies of micromachined hot-wire anemometers,” *Technical Digest, International Electron Devices Meeting 1994,* San Francisco, CA, pp. 139–142, December 1994.
12. Ebefors, T., Kalvesten, E., and Stemme, G., “Three dimensional silicon triple-hot-wire anomometer based on polyimide joints,” *Proceedings of the 11*th *International Workshop on MEMS,* pp. 93–98, January 1998.

20 Actuators

George T.-C. Chiu
Purdue University

C. J. Fraser
University of Abertay Dundee

Ramutis Bansevicius
Kaunas University of Technology

Rymantas Tadas Tolocka
Kaunas University of Technology

Massimo Sorli
Politecnico di Torino

Stefano Pastorelli
Politecnico di Torino

Sergey Edward Lyshevski
Purdue University Indianapolis

20.1 Electromechanical Actuators

George T.-C. Chiu

Introduction

As summarized in the previous sections, a mechatronics system can be partitioned into function blocks illustrated in Fig. 20.1. In this chapter, we will focus on the actuator portion of the system. Specifically, we will present a general discussion of the types of electromechanical actuators and their interaction

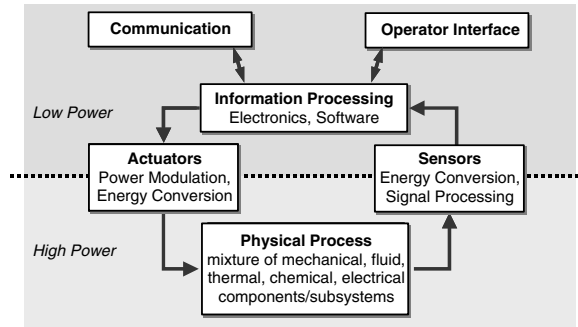



FIGURE 20.1 Mechatronic system.

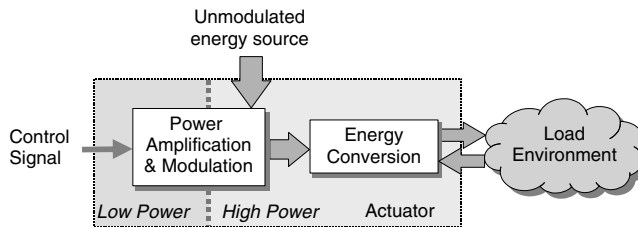



FIGURE 20.2 Actuator functional diagram.

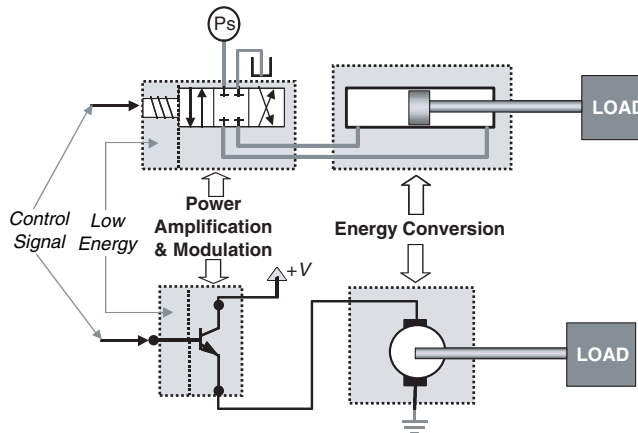



FIGURE 20.3 Electrohydraulic and electromechanical actuators.

with the load or physical environment. In addition, we will look at the electronic components that are essential for modulating the necessary electrical energy.

From an energy perspective, a mechatronic system can be separated into a relatively higher energy (power) portion that interacts with the physical world and a relatively low energy (power) portion that process the data, see Fig. 20.1. Sensor and actuators are the interfacing devices that accomplish the tasks of energy modulation and energy conversion. Therefore, an actuator can be viewed as having the structure depicted in Fig. 20.2. Typically, actuators are considered only as energy conversion devices. However, with the proliferation of power electronics, we will take a more inclusive view of actuators that also includes power amplification. An electrohydraulic linear actuator, see Fig. 20.3, can also be similarly classified, where the spool valve is the power amplification/modulation block with spool position as the

control signal and the hydraulic pressure/flow is the energy source. The hydraulic cylinder acts as the energy conversion device that converts fluidic energy to mechanical energy. For a typical electromechanical actuator, such as a DC motor (Fig. 20.3), the power amplification block is the motor driver that amplifies signal level (low current/power) control signal to the higher power (large current) signal that is used to convert electrical energy to mechanical energy through the electromagnetic principle.

In this chapter, we will first present an overview of common types of electromechanical actuators. They will be classified by the respective energy conversion mechanism. The power electronic components, such as diodes, thyristors, and transistors, which are used for power amplification and modulation, will be presented followed by discussion of common power amplification building blocks. We will conclude by discussing some issues related to interfacing with electromechanical actuators.

Type of Electromechanical Actuators—Operating Principles

Converting electrical energy to mechanical energy is the common thread among different electromechanical actuators. Physics provided us with many different mechanisms either through direct conversion such as piezoelectric or through an intermediate medium such as a magnetic field. We will present an overview of the more common electromechanical actuators by their energy conversion mechanism: electromagnetic, electrostatic, and piezoelectric. The following discussion is intended to provide introductory information about the types of electromechanical actuation and is by no means exhaustive. Detailed discussion of each can be found in subsequent chapters, where they will be discussed in more detail.

Electromagnetics—Magnetic Field

Electromagnetic is the most widely utilized method of energy conversion for electromechanical actuators. One of the reasons for using magnetic fields instead of electric fields is the higher energy density in magnetic fields. The air gap that separates a stationary member (stator) and a moving member of an electromechanical actuator is where the electromechanical energy conversion takes place. The amount of energy per unit volume of air gap for magnetic fields can be five orders of magnitude higher than that of electric fields.

Lorentz's law of electromagnetic forces and Faraday's law of electromagnetic induction are the two fundamental principles that govern electromagnetic actuators. Before going into the detail of electromagnetics, we will first introduce the concept of magnetic field and flux.

Magnetic flux ϕ exists due the presence of a magnetic field. The magnetic field strength $\vec{H}$ (in A/m) and the magnetic flux density $\vec{B}$ (in tesla [T]) are related by the permeability of the material. In a vacuum, the magnetic flux density is directly proportional to the magnetic field strength and is expressed by

$$\vec{B} = \mu_0 \cdot \vec{H} \tag{20.1}$$

where $\mu_0 = 4$[T m/A] is the *permeability constant*. For other magnetic or ferromagnetic materials the relationship is given by

$$\vec{B} = \mu_r(\vec{H}) \cdot \mu_0 \cdot \vec{H} \tag{20.2}$$

where $\mu_r(\vec{H})$ is the *relative permeability* of the material. Figure 20.4 shows typical *B*-*H* and μ-*H* curves.

Lorentz's Law of Electromagnetic Force

When a current carrying conductor is placed in a magnetic field, it will be subjected to an induced force given by

$$\vec{F} = \vec{i} \times \vec{B} \tag{20.3}$$

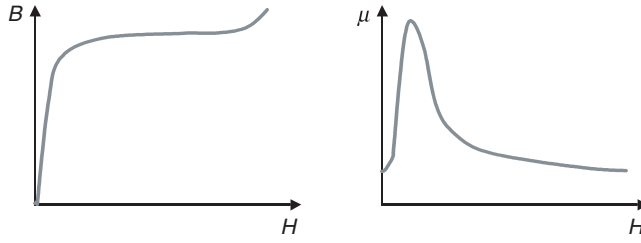



FIGURE 20.4 μ-H diagram and B-H diagram.

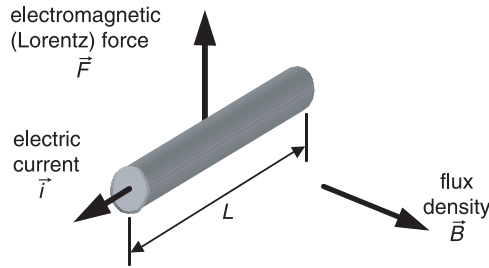



FIGURE 20.5 Lorentz's electromagnetic force.

where $\vec{F}$ is the force vector, $\vec{i}$ is the current vector, and $\vec{B}$ is the magnetic flux density. The force is called the *electromagnetic force* or the Lorentz force. If a conductor of length L carrying constant current i is placed in a constant (independent of location) field B, as shown in Fig. 20.5, the magnitude of the resultant Lorentz force $\vec{F}$ exerted by the field B on the conductor is

$$F = |\vec{F}| = BLi \tag{20.4}$$

Faraday's Law of Electromagnetic Induction

The motion of a conductor in a magnetic field will produce an electromotive force (emf), or electric potential, across the conductor given by

$$\text{emf} = E = -\frac{d\phi}{dt} \tag{20.5}$$

where $\phi = \oint \vec{B} \cdot d\vec{A}$ is the magnetic flux. For a conductor of length L moving at a constant speed v in a constant (independent of location) magnetic field that is perpendicular to the area A, as shown in Fig. 20.6, the magnitude of the induced electromotive force (electric potential) is

$$\text{emf} = E = BLv \tag{20.6}$$

There are two methods to generate a desired magnetic field $\vec{H}$, or equivalently, a desired magnetic flux density $\vec{B}$. One is to use a permanent magnet and the other is to utilize the Boit–Savart law.

Boit–Savart Law

A long (infinite), straight, current carrying conductor induces a magnetic field around the conductor, see Fig. 20.7. The flux density at a perpendicular distance r from the conductor is

$$B = \frac{\mu_r \mu_0}{2\pi r} \cdot i \tag{20.7}$$

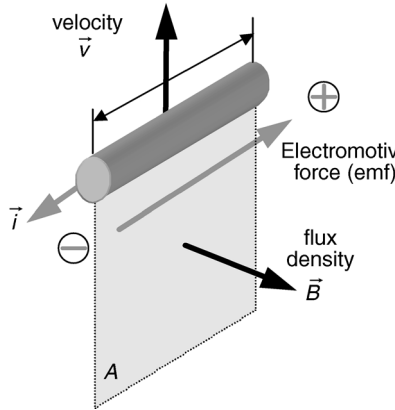



FIGURE 20.6 Motion induced electromotive force.

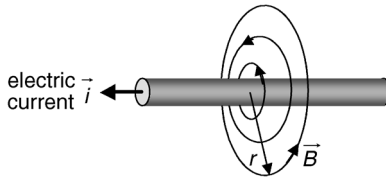



FIGURE 20.7 Magnetic field generated by current carrying conductor.

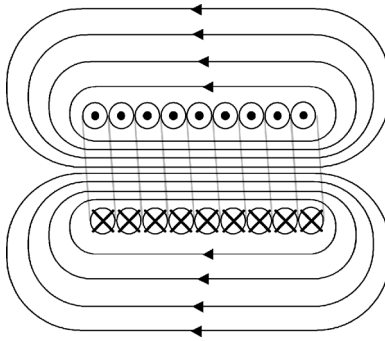



FIGURE 20.8 Coil (solenoid) induced magnetic field.

where i is the electric current. If we bend the straight current carrying conductor into a helical coil (solenoid) with N turns, it will induce a corresponding magnetic field as depicted in Fig. 20.8. If the length of the coil L is much greater than its diameter D, the flux density follows the right-hand rule and the magnitude inside the coil is approximately

$$B = \mu \frac{N}{L} \cdot i \tag{20.8}$$

where $\mu = \mu_r \cdot \mu_0$ is the permeability of the material inside the coil and i is the current through the winding. This field can be intensified by inserting a ferromagnetic core into the solenoid by increasing the permeability. Coil induced magnetic fields is widely utilized in electromagnetic devices for generating controlled magnetic fields and are often referred to as *electromagnets*.

FIGURE 20.9 Assorts of solenoid actuators. (Courtesy of Shih Hsing Industrial Co., Ltd.)

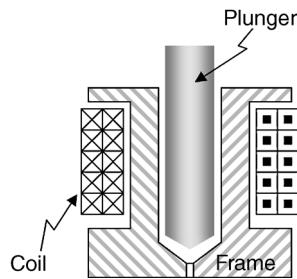



FIGURE 20.10 A typical solenoid.

Solenoid Type Devices

Solenoids, see Fig. 20.9, is the simplest electromagnetic actuators that are used in linear as well as rotary actuations for valves, switches, and relays. As the name indicates, a solenoid consists of a stationary iron frame (stator), a coil (solenoid), and a ferromagnetic plunger (armature) in the center of the coil, see Fig. 20.10.

As the coil is energized, a magnetic field is induced inside the coil. The movable plunger moves to increase the flux linkage by closing the air gap between the plunger and the stationary frame. The magnetic force generated is approximately proportional to the square of the applied current i and is inverse proportional to the square of the air gap δ, which is the stroke of the solenoid, i.e.,

$$F \propto \frac{i^2}{\delta^2} \tag{20.9}$$

As shown in Fig. 20.11, for strokes less than 0.060 in., the flat face plunger is recommended with a pull or push force three to five times greater than 60° plungers. For longer strokes up to 0.750 in., the 60° plunger offers the greatest advantage over the flat face plunger. When the coil is de-energized, the field decreases and the plunger will return to the original location either by the load itself or through a return spring.

All linear solenoids basically pull the plunger into the coil when energized. Push-type solenoids are implemented by extending the plunger through a hole in the back-stop, see Fig. 20.12. Therefore, when energized, the plunger is still pulled into the coil, but the extended producing a pushing motion from the back end of the solenoid. Return motion, upon de-energizing the coil, is provided by the load itself (i.e., the weight of the load) and/or by a return spring, which can be provided as an integral part of the solenoid assembly.

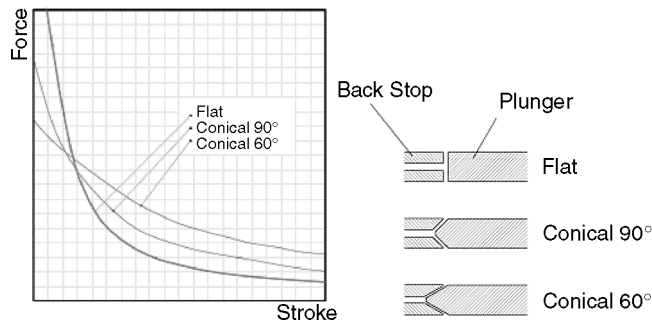



FIGURE 20.11 Typical force-stroke curve of solenoids. (Courtesy of Magnetic Sensor Systems.)

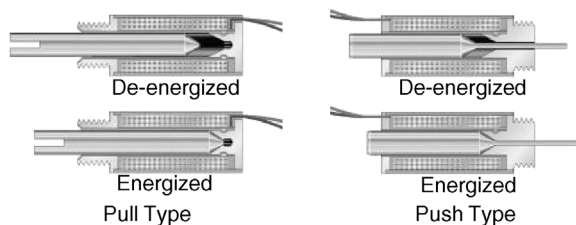



FIGURE 20.12 Push and pull type solenoids. (Courtesy of Ledex® & Dormeyer® Products.)

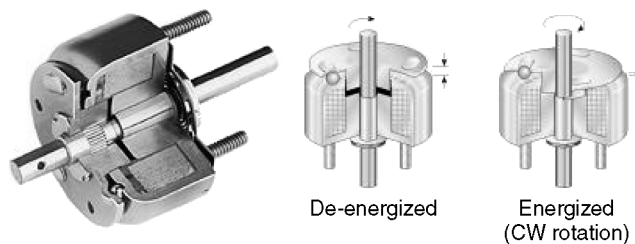



FIGURE 20.13 Rotary solenoid. (Courtesy of Ledex® & Dormeyer® Products.)

Rotary solenoids utilize ball bearings that travel down inclined raceways to convert linear motion to rotary motion. When the coil is energized, the plunger assembly is pulled towards the stator and rotated through an arc determined by the coining of the raceways, see Fig. 20.13. An *electromechanical relay* (*EMR*) is a device that utilizes a solenoid to close or open a mechanical contact (switch) between high power electrical leads. A relay performs the same function as a power transistor in that relatively small electrical energy is used to switch a large amount of currents. The difference is that a relay has the capability of controlling much larger current level. Variations on this mechanism are possible: some relays have multiple contacts, some are encapsulated, some have built-in circuits that delay contact closure after actuation, and some, as in early telephone circuits, advance through a series of positions step by step, as they are energized and de-energized.

Design/Selection Considerations. Force, stroke, temperature, and duty cycle are the four major design/selection considerations for solenoids. A linear solenoid can provide up to 30 lb of force from a unit less than 2¼ in. long. A rotary solenoid can provide well over 100 lb of torque from a unit also less than 2¼ in. long. As shown in Fig. 20.11, the relationship between force and stroke can be modified by changing the design of some internal components. Higher performance, e.g., force output, can be

TABLE 20.1 Temprature Rating For Electrical Insulations

Insulation Classification		Temperature Rating	
Class A	Class 105	105°C	221°F
Class E	Class 120	120°C	248°F
Class B	Class 130	130°C	266°F
Class F	Class 155	155°C	311°F
Class H	Class 180	180°C	356°F
Class N	Class 200	200°C	392°F

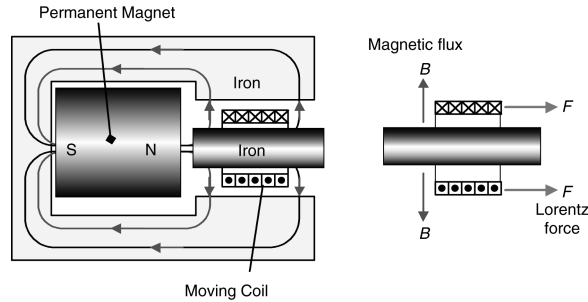



FIGURE 20.14 Voice-coil motor.

achieved by increasing the current to the coil winding. However, higher current tends to increase the winding temperature. As the winding temperature increases, the wire resistance increases. This will reduce the output force level. Solenoids are often rated as operating under continuous duty cycle or intermittent duty cycle. A solenoid rated for 100% duty cycle may be energized at its rated voltage continuously because its total coil temperature will not exceed maximum allowable ratings, while an intermittent duty cycle solenoid has an associated allowable "on" time which must not be exceeded. Intermittent duty coils provide considerably higher forces than continuous duty solenoids. The maximum operating temperature for a solenoid is determined by the rated temperature of the insulation material used in the winding (see Table 20.1).

Voice-Coil Motors (VCMs)

As the name indicates, the voice-coil motor was originally developed for loudspeakers. It is now extensively used in moving read/write heads in hard disk drives. Since the coil is in motion, VCM is also referred to as a *moving-coil* actuator. The VCM consists of a moving coil (armature) in a gap and a permanent magnet (stator) that provides the magnetic field in the gap, see Fig. 20.14. When current flows through the coil, based on the Lorentz law, the coil experiences electromagnetic (Lorentz) force F

$$\vec{F} = \vec{i} \times \vec{B}$$

Since most voice-coils are designed so that the flux is perpendicular to the current direction, the resultant Lorentz force can be written as

$$F_{\mathrm{VCM}} = \gamma BNl \cdot i = K_F \cdot i \Rightarrow F_{\mathrm{VCM}} \propto i \tag{20.10}$$

where l is the coil length per turn, B is the flux density, N is the number of turns in the coil, i is the current, and γ is a coil utilization factor. It is important to know that the force is proportional to the applied current amplitude and the proportional constant K_F is often called the *force constant.*

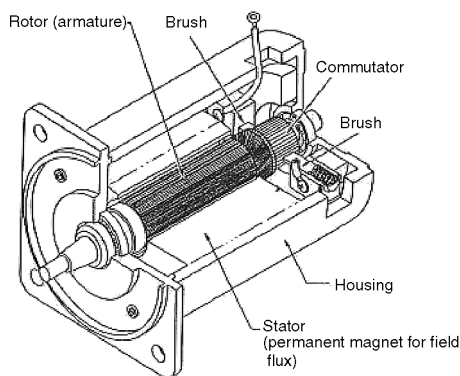



FIGURE 20.15 Permanent magnet DC motor. (T. Keujo and S. Nagamori, *Permanent-Magnet and Brushless DC Motors,* 1985, by permission of Oxford University Press.)

The coil is usually suspended in the gap by springs and attached to the load such as the diaphragm of an audio speaker, the spool of a hydraulic valve, or the read/write head of the disk drive. The linear relationship between the output force and the applied current and the bidirectional capability makes the voice coil more attractive than solenoids. However, since the controlled output of the voice coil is force, some type of closed loop control or some type of spring suspension is needed.

Design/Selection Consideration. From Eq. (20.10) we see that the force constant depends on the flux density and the amount of wires that can be packed into the gap. There are two options to increase the force constant. One is to increase the flux density, which can be achieved by using stronger magnetic material and the other is to increase either N or l, i.e., to pack more turns and/or make a larger diameter coil.

Given a fixed gap volume, using higher gauge (thinner) wires is the only way to increase the number of turns. However, higher gauge wires have larger resistance, which will increase the resistive heating of the winding and limit the allowable current. In addition, the additional insulation will also occupy more volume and tends to reduce the effect of increasing N. In summary, to improve the performance of the voice coil, a designer can either choose a better magnetic material or to make the motor bigger by either making the coil wider (increase D) or longer (increase N).

Electric Motors

Electric motors are the most widely used electromechanical actuators. They can either be classified based on functionality or electromagnetic characteristics. The differences in electric motors are mainly in the rotor design and the method of generating the magnetic field. Figure 20.17 shows the composition of a permanent magnet DC motor. Some common terminologies for electric motors are:

Stator is the stationary outer or inner housing of the motor that supports the material that generates the appropriate stator magnetic field. It can be made of permanent magnet or coil windings.

Field coil (*system*) is the portion of the stator that is responsible for generating the stator (field) magnetic flux.

Rotor is the rotating part of the motor. Depending on the construction, it can be a permanent magnet or a ferromagnetic core with coil windings (armature) to provide the appropriate armature field to interact with the stator field to create the torque.

Armature is the rotor winding that carries current and induces a rotor magnetic field.

Air gap is the small gap between the rotor and the stator, where the two magnetic fields interact and generate the output torque.

Brush is the part of a DC motor through which the current is supplied to the armature (rotor). For synchronous AC motors, this is done by *slip rings*.

Commutator is the part of the DC motor rotor that is in contact with the brushes and is used for controlling the armature current direction. Commutation can be interpreted as the method to control the current directions in the stator and/or the armature coils so that a desired relative stator and rotor magnetic flux direction is maintained. For AC motors, commutation is done by the AC applied current as well as the design of the winding geometry. For stepping motors and brushless DC (BLDC) motors, commutations are done in the drive electronics and/or motor commands.

Torque generation in an electric motor is either through the interaction of the armature current and the stator magnetic field (Lorentz Law) or through the interaction of the stator field and the armature field. Table 20.2 summarizes the common classification of electric motors. The next chapter will give a detailed discussion of the operation of various electric motors and the associated design considerations.

TABLE 20.2 Electric Motor Classification

Classification			
Command Input	Magnetic Field		Description
DC motors	Permanent magnet		Permanent magnets are used to generate the stator magnetic field. Electrical current is supplied directly into the armature winding of the rotor through the brushes and commutators.
	Electro-magnets	Shunt wound	A stator (field) winding is used as electromagnet. Stator winding is connected in parallel with the armature winding.
		Series wound	A stator (field) winding is used as electromagnet. Stator winding is connected in series with the armature winding.
		Compound wound	Two stator (field) windings are used as electromagnet. The stator windings are connected, one in series and one in parallel, with the armature winding.
		Separate wound	A stator (field) winding is used as electromagnet. Both the stator and armature fields are individually energized.
AC motors	Single-phase	Induction	Single stator winding with squirrel-cage rotor. No external connection to the rotor. Torque generation is based on the electromagnetic induction between the stator and rotor. AC current provides the commutation of the fields. Rotor speed is slightly slower than the rotating stator field (slip).
		Synchronous	Permanent magnet rotor or rotor winding with slip ring commutation. Rotating speed is synchronized with the frequency of the AC source.
	Poly-phase	Induction	Similar to single-phase induction motor but with multiple stator windings. Self-starting.
		Synchronous	Similar to single-phase synchronous motor but with multiple stator windings for smoother operation.
	Universal		Essentially a single-phase AC induction motor with similar electrical connection as a *series wound* DC motor. Can be driven by either AC or DC source.
Stepper Motors	Permanent magnet		Permanent magnet rotor with stator windings to provide matching magnetic field. By applying different sequence (polarity) of coil current, the rotor PM field will align to match induced stator field.
	Variable reluctance		Teethed ferromagnetic rotor with stator windings. Rotor motion is the result of the minimization of the magnetic reluctance between the rotor and stator poles.
	Hybrid		Multi-toothed rotor with stator winding. The rotor consists of two identical teethed ferromagnetic armatures sandwiching a permanent magnetic.
Brushless DC motors	Poly-phase	Synchronous	Essentially a poly-phased AC synchronous motor but using electronic commutation to match rotor and stator magnetic fields. Electronic commutation enables using a DC source to drive the synchronous motor.

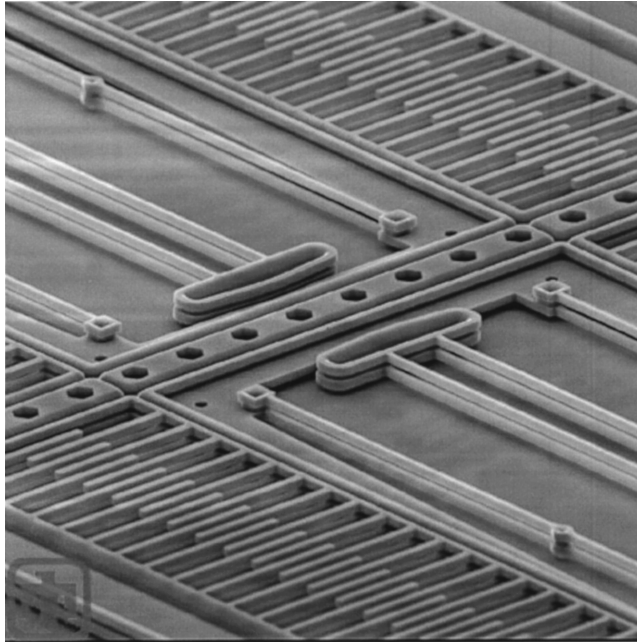



FIGURE 20.16 MEMS comb actuator uses electrostatic actuation. (Courtesy of Sandia National Laboratories, MEMS and Novel Si Science and Technology Department, SUMMIT Technologies, www.mems.sandia.com.)

Electrostatics—Electrical Field

Since electrical fields have lower energy density than magnetic fields, typical applications of electrical field forces are limited to measurement devices and accelerating charge particles, where the required energy density is small. Recently, with the proliferation of microfabrication technology, it is possible to apply the small electrostatic forces to microelectromechanical actuators, such as comb actuators (see Fig. 20.16). The advantage of electrostatic actuation is the higher switching rate and less energy loss as compared to the electromagnetic actuation. However, the limitation in force, travel, and high operating voltage still needs to be addressed. Electrostatic actuation is the main actuation for moving charged toner particles in electrophotographic (xerographic) processes, e.g., laser printers.

Piezoelectric

Piezoelectric is the property of certain crystals that produces a voltage when subjected to mechanical deformation, or undergoes mechanical deformation when subjected to a voltage. When a piezoelectric material is under mechanical stress, it produces an asymmetric displacement in the crystal structure and in the charge center of the affected crystal ions. The result is charge separation. An electric potential proportional to the mechanical strain can be measured. This is called the *direct piezoelectric effect*. Conversely, the material will have deformation without volume change when electric potential is applied. This *reciprocal piezoelectric effect* can be used to produce mechanical actuation. There are two categories of piezoelectric materials: sintered ceramics, such as lead-zirconate-titinate (PZT), and polymers, such as polyvinylidence fluoride (PVDF). Piezoceramics have a larger force output and are used more as actuators. PVDFs tend to generate larger deformation and are used more for sensor applications.

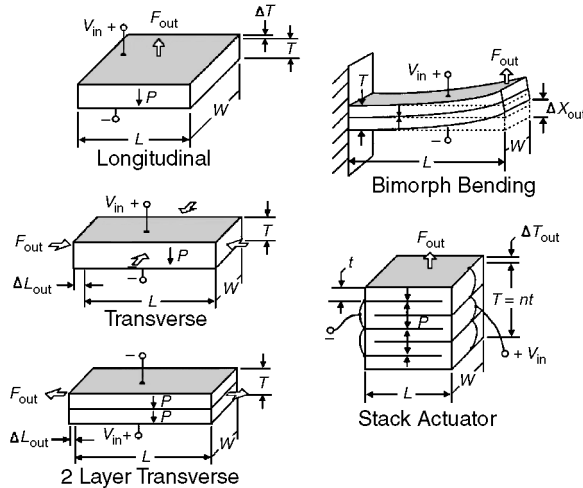



FIGURE 20.17 Common piezoelectric actuation geometries. (Courtesy of Piezo Systems, Inc.)

The coupling between the electrical and mechanical property of the material can be modeled by the following sets of linear constitutive equations:

$$\begin{cases} S = s^E \cdot T + d \cdot E \\ D = d \cdot T + \varepsilon^T \cdot E \end{cases} \Leftrightarrow \begin{cases} E = -g \cdot T + (\varepsilon^T)^{-1} \cdot D \\ S = s^D \cdot T + g \cdot D \end{cases} \tag{20.11}$$

where

E = electric field strength [V/m]
D = charge-density (dielectric) displacement [C/m^2]
T = stress [N/m^2]
S = strain
s = compliance
ε = permittivity [F/m]
d, g = piezoelectric coupling coefficients

Design/Selection Considerations

Figure 20.17 shows the common orientation for piezoelectric actuation. With a typical strain of less than 0.3%, the amount of deflection or deformation is usually the limited factor for piezoelectric actuators. The most common architectures are the stacked and bending actuation. Piezoelectric actuators are most suited for high bandwidth, large force, and small stroke/deflection applications. They are widely used in noise and acoustical applications, as well as optical applications, where precision motion is critical.

Piezoelectric actuators are usually specified in terms of their free deflection and blocked force. Free deflection (X_f) refers to displacement attained at the maximum recommended voltage level when the actuator is completely free to move and is not asked to exert any force. Blocked force (F_b) refers to the force exerted at the maximum recommended voltage level when the actuator is totally blocked and not allowed to move. Figure 20.18 shows the static performance curve of a typical piezoelectric actuator (force vs. deflection). Generally, a piezo actuator must deform a specified amount and exert a specified force, which determines its operating point on the force vs. deflection line. An actuator is considered optimized for a particular

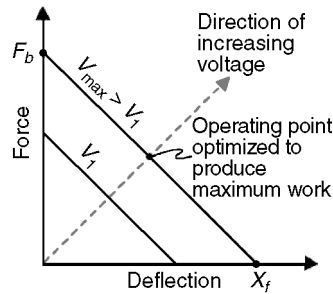



FIGURE 20.18 Static performance curve of a typical piezoelectric actuator. (Courtesy of Piezo Systems, Inc.)

application if it delivers the required force at one half its free deflection. High operating voltage, hysteresis, creep, and fatigue are the main mechanical design considerations.

Efficiency

Efficiency is one of the major considerations for any energy conversion process. In most cases, the wasted energy is converted to heat and increases the device temperature. For electromechanical actuators, heat (temperature) is one of the most prominent performance-limiting factor as well as failure mode. As device temperature increases, the underlying conversion efficiency will suffer and dump more energy into heat, which further increases the device temperature. This is often referred to as *thermal runaway*. Therefore, it is very important when designing electromechanical actuators to prevent thermal runaway and guarantee that under normal operating condition the actuator system achieves thermal equilibrium. The equilibrium temperature should be maintained below the lowest rated temperature of the components, such as the electrical insulation for the windings. The temperature rating for electrical insulations are listed in Table 20.1.

Power Amplification and Modulation—Switching Power Electronics

As described in the previous section and depicted in Fig. 20.2, there are two main functions in an extended definition of an actuator for mechatronics systems. We have introduced a few energy conversion mechanisms and the associated actuators. In the second part of this chapter, we will focus on the power amplification and modulation portion of the actuator. This part of the actuator is traditionally called the *power amplifier* or the *driver* for the corresponding actuator. However, as miniaturization and system integration become more pervasive, power electrics are being embedded into either the controller (information processing unit) or the actuator. It is also the portion where intelligence and additional functionality/feature can be incorporated. For electromechanical actuators, the unmodulated energy source is electricity. The power amplifier acts as a buffer between the low energy part of the system, where actuation command is given in low energy electrical signals, and the high energy density electrical signal that will be converted.

Power amplification can roughly be categorized into two methods, linear and switching. The main advantage of linear power amplification is the "cleanness" of the signal as compared to the switching amplifiers. The main drawback is in efficiency, where linear amplifiers tend to run hotter than similar sized switching amplifiers. However, as with any engineering design, this is only a rule-of-thumb; the designer needs to analyze the application and select or design the appropriate driver.

Switching amplifiers are made of semiconductor components such as diodes and transistors. These semiconductor devices either function as a switching element that controls the current flow to the energy conversion element such as a winding coil, or as an amplification element that modulates the amount of current flowing into the winding coil. Another advantage of using switching type power amplifiers is that, with switching, the amplifier stage can be directly controlled by a digital signal from an information processing device (see Fig. 20.1) such as a microcontroller or a microprocessor. This eliminates the need

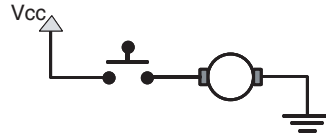



FIGURE 20.19 DC motor under switching control.

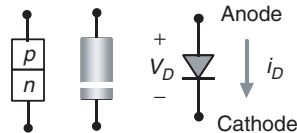



FIGURE 20.20 Diode.

for an ADC, which reduces the cost and size of the required electronics. Pulse-width modulation (PWM) is one good example using binary signal to control electromechanical actuators. Figure 20.19 shows one simple example of using a switching device to interface with a PMDC motor. As discussed in the previous section, torque is generated when current is flowing into the armature winding of a PMDC motor. To turn the motor on and off, we can connect the motor with a DC power source in series with a switch. When the switch is closed, current flows through the motor and the motor turns. If the switch is opened, current stops and the motor will eventually stop. Of course, more sophisticated circuit and switching design is needed for actual implementation.

Nevertheless, this example illustrates the fundamentals of switching power amplification. In this section, we will focus on switching amplifiers by introducing the fundamental building blocks.

Semiconductors

Semiconductors are typically materials consisting of elements from group IV of the periodic table, e.g., silicon (Si), germanium (Ge), and cadmium sulfide. Unlike conductors and insulators, semiconductors' current-carrying capability is significantly affected by the temperature and the amount of incident photons and the type and amount of impurities in the material. By introducing carefully controlled group V or III elements (called *dopants*) into the semiconductors, we can increase or decrease the number of valence electrons in the semiconductors, respectively. Depending on the type of the dopants, semiconductors can be separated into:

- *n-type semiconductors*: semiconductors doped with *donor* elements (e.g., arsenic or phosphor group V elements) that result in one additional electron freed (*free electron*) from the crystal lattice as a charge carrier that is available for conducting.
- *p-type semiconductors*: semiconductors doped with *acceptor* elements (e.g., boron or gallium group III elements) that results in a missing electron in the lattice structure, which is called a *hole*. Holes can be viewed as positive charge carriers or places that accept free electrons.

As will be discussed shortly, the interaction between the n-type and p-type semiconductors under different orientation forms the basis for all the semiconductor electronic devices. One of the more interesting aspects of modern electronics is the variety of features that can be obtained with a simple switching device that opens or closes a connection in a controlled manner. We will discuss a few electronic elements that are widely used, mainly as a controlled switching element, in power electronics for constructing power amplifiers/drivers for electromechanical actuators.

Diodes

A diode is a two-terminal electronic device that is constructed by joining a p-type and an n-type semiconductor together to form a *pn junction*. Figure 20.20 shows the schematic symbol of a generic diode. The terminal associated with the p-type material is called the *anode* and the terminal associated with the n-type material is called the *cathode*. If the anode has higher electrical potential (>0.7 V) than

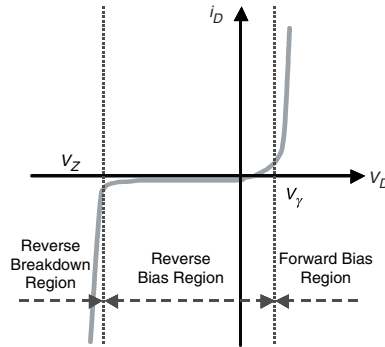



FIGURE 20.21 Diode characteristics.

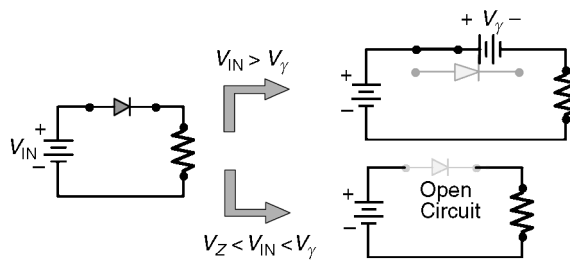



FIGURE 20.22 Approximating a diode in a circuit.

the cathode, the diode is said to be *forward biased*, i.e., $V_D > 0.7$ V. Conversely, if $V_D < 0.7$ V, the diode is *reverse biased*.

As shown in Fig. 20.21, depending on the applied voltage, a diode can operate in three different regions:

- *Forward biased region*: $V_D > V_\gamma$, where V_γ is called the *forward bias voltage* and is typically around 0.7 V for silicon and 0.3 V for germanium. The diode acts as a closed switch, and the anode and cathode become short-circuited with a slight reverse potential (forward voltage drop) that is equal to V_γ, see Fig. 20.22.
- *Reverse biased region*: $V_Z < V_D < V_\gamma$, where V_Z is the *reverse breakdown voltage* of the diode. The diode acts as an opened switch and the circuit is open, see Fig. 20.22.
- *Breakdown region*: $V_D < V_Z$. The diode again acts as a closed switch and a large current flows through the diode. This is called the *avalanche* effect. If the magnitude of the reverse current i_D is larger than the critical reverse bias current, the device will fail.

A diode is analogous to a fluid check valve, which allows fluid (current) to flow in only one direction if the forward pressure is sufficient to overcome the spring force, see Fig. 20.23. Table 20.3 summarizes properties of some typical diodes.

Maximum allowable current through a diode and the reverse breakdown voltage are the two major design considerations for diodes. The voltage across a diode times the current it carries is the power loss across the diode that is completely converted into heat. The temperature of a diode can rise rapidly due to its small size and mass. For safe operation, the temperature of the diode junction should not exceed 200°C. To improve heat transfer, diodes are commonly mounted on metallic *heat sinks*. Signal diodes are rated between $^1/_2$ and 1 W. Power diodes can be rated as high as several hundred kilowatts.

TABLE 20.3 Typical Diode Properties

Relative Power	Maximum Average Forward Current [A]	Voltage Drop at Maximum Average Forward Current [V]	Maximum Peak Forward Current [A]	Reverse Breakdown Voltage [V]	Maximum Junction Temperature [°C]
Low	1	0.8	30	1000	175
Medium	12	0.6	240	1000	200
High	100	0.6	1600	1000	200
Very High	1000	1.1	10,000	2000	200

Source: T. Wildi, *Electrial Machines, Drivers and Power Systems,* Prentice Hall, 2000.

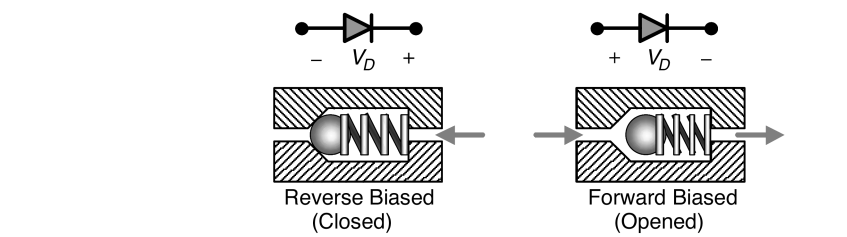



FIGURE 20.23 Analogy between diodes and check valves.

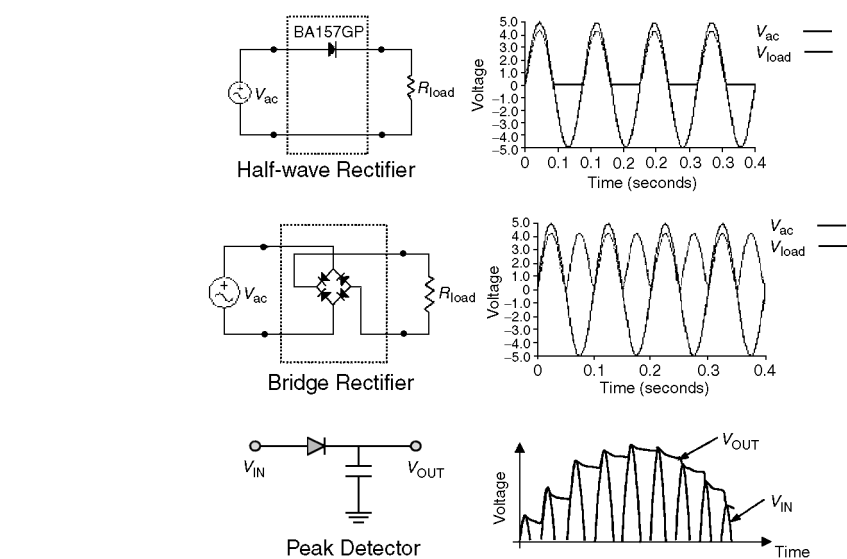



FIGURE 20.24 Common diode applications.

For switching applications, the *reverse-recovery time* is another important design parameter. The reverse-recovery time imposes an upper bound on the frequency at which the diode can be switched on and off. Attempts to operate a diode above this frequency will result in a decrease in switching efficiency and may cause severe overheating.

Diodes are widely used in electronic power circuits. They are most widely used for rectification and peak detection. Figure 20.24 illustrates some of the common diode applications. If multiple diodes are to be

used, diode array can be used. In general, the term diode array implies four or more diodes in a single package. The most efficient packaging scheme is typically eight diodes or more in a dual inline package (DIP). Other packages are the single inline package (SIP), the flat pack, and even a surface mount diode array. Although multiple diode arrays can incorporate different type diodes, the most popular arrays incorporate a fast, small signal diode such as the 1N4148, and the core driver arrays, which employ a fast switching, higher current, 100-mA diode.

Zener Diode

Recall the current–voltage curve of a diode shown in Fig. 20.21. If a diode is reverse biased to the breakdown region, a large reverse current will flow through the diode. For most diodes, this voltage is usually larger than 50 V and may exceed kilovolts. *Zener* (*Avalanche*) diodes are a class of diodes that exhibit a steep breakdown curve with a well-defined reverse breakdown voltage V_Z. This unique breakdown characteristic makes Zener diodes good candidates for building *voltage regulators*, since they can maintain a stable source voltage under variable supply as well as varying load impedance. Figure 20.25 shows the special symbol that represents a Zener diode.

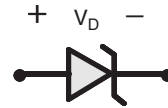



FIGURE 20.25 Zener (avalanche) diode.

To use the Zener diode as a voltage regulator, it should be reverse biased with a supply voltage higher than the rated reverse breakdown voltage V_Z, see Fig. 20.26. For an ideal Zener diode, in Fig. 20.26, $V_S > V_Z$, the voltage across the load will equal to V_Z; hence the load current i_{load} and the Zener diode current can be written as

$$i_{load} = \frac{V_Z}{R_{load}} \quad \text{and} \quad i_Z = \frac{V_S - V_Z}{R_S} \quad .$$

Zener diodes are often rated by their power dissipation, which is

$$P_{Zmax} = i_{Zmax} \cdot V_Z \quad .$$

Therefore, when selecting Zener diode for voltage regulation applications, it is important to ensure that i_{Zmax} does not exceed the allowable limit. The most common range of the reverse breakdown voltage for Zener diodes is from 3.3 V to 75 V. However, voltages out of this range are available. Some typical power ratings for Zener diodes are 1/4, 1/2, 1, 5, 10, and 50 W.

Thyristors

A thyristor, or a *silicon-controlled rectifier* (*SCR*), is a 4-layer semiconductor switch, similar to a diode, but with an additional terminal to control the instant of conduction. A thyristor has three terminals: an anode, a cathode, and a gate, see Fig. 20.27. One can think of a thyristor as a controllable diode that the gate terminal provides as a mean of precise control of the instance when the thyristor is to be turned on, i.e., it is a controlled switch.

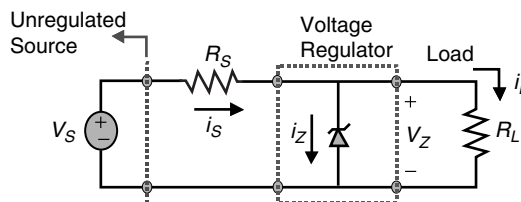



FIGURE 20.26 Use zener diode as simple voltage regulator.

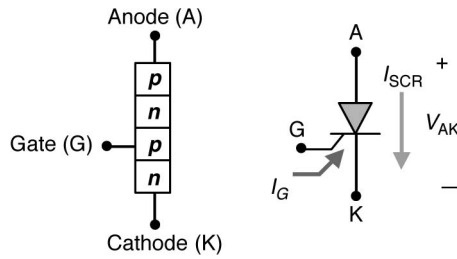



FIGURE 20.27 Thyristor and its schematic symbol.

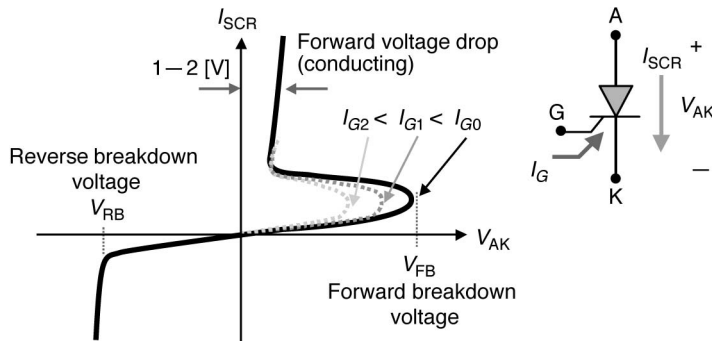



FIGURE 20.28 Thyristor (SCR) characteristics.

Figure 20.28 shows the current–voltage characteristics of a thyristor. To turn on (short or conduct) a thyristor, two conditions have to be satisfied:

1. The anode (A) and cathode (K) terminal has to be forward biased, i.e., the anode voltage needs to be higher than the cathode voltage.
2. A gate current I_G has to flow into the gate for a sufficient amount of time, typically, a few microseconds. The gate current can be generated by a short positive voltage pulse applied across the gate (G) and cathode (K) terminal. The minimum amount of gate current that is required to turn on a thyristor is called the *latching current*.

When the thyristor is turned on, the amount of current flowing through the device I_{SCR} is limited by the rest of the circuit impedance. Once the thyristor is turned on, the gate terminal loses control of the device, i.e., we cannot use the gate to turn off the device. The thyristor will only turn off if the anode current I_{SCR} goes to zero, after which the gate terminal can assert control to turn on the device again. Obviously, the thyristor can also be switched on by exceeding the forward breakdown voltage V_{FB}. However, this is usually considered a design limitation and switching is normally controlled with a gate voltage. If the gate (G) terminal is shorted with the cathode (K), the thyristor cannot be turned on, even if V_{AK} is forward biased. One can think of the thyristor as a normally opened switch with a detent. Once the switch is closed, no additional control is needed. Figure 20.29 shows the operation of a thyristor driving a simple resistive load under a sinusoidal bipolar source voltage V_S. In Fig. 20.31, the gate voltage V_G will be the command or control input.

When the thyristor is reverse biased, the gate (G) to cathode (K) terminals should not be forward biased to prevent reverse breakdown of the first pn junction of the thyristor, see Fig. 20.27. The reverse breakdown voltage V_{RB}, the latching current, the current and power rating, and the rate of rise of voltage are the more important design parameters for selecting a thyristor. When the voltage across the thyristor is suddenly applied or increased rapidly, the thyristor may turn on even if the gate current (voltage) is

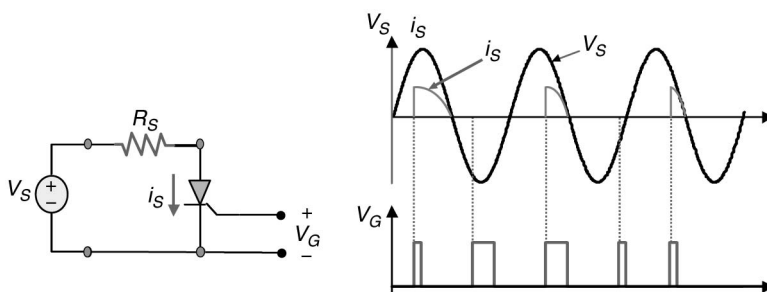



FIGURE 20.29 Thyristor driving a resistive load.

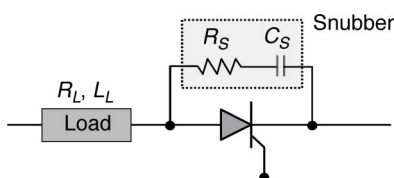



FIGURE 20.30 Snubber circuit.

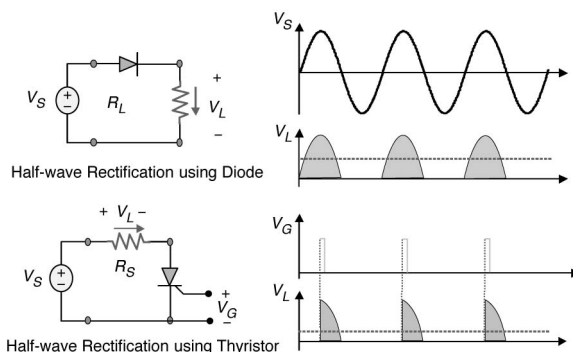



FIGURE 20.31 Controlled (thyristor) and uncontrolled (diode) rectifications.

zero. A typical rate of voltage change that will induce thyristor turn-on is about 50 V/μs. To prevent undesired conduction due to a large rate (high frequency) of voltage variation, a *snubber circuit*, see Fig. 20.30, is often connected in parallel with the thyristor to filter out the high frequency voltage variations. The snubber circuit is essentially a passive RC low-pass filter. The selection of the snubber resistance R_S and capacitance C_S can use the following formula:

$$C_S = \frac{V_{A\max}^2}{L_L(dV/dt)_{\max}^2} \quad \text{and} \quad R_S = 2\sqrt{\frac{L_L}{C_S}} - R_L \tag{20.12}$$

where R_L and L_L are the load inductance and load resistance, respectively. $V_{A\max}$ is the maximum anode voltage and $(dV/dt)_{\max}$ is the maximum expected rate of raise of voltage across the anode and cathode.

Unlike diodes used in rectifier circuits that can only rectify the input AC voltage to an average DC voltage, thyristors can be used to build controlled rectifiers that can rectify AC sources and modulate the average output DC voltage by modulating the firing timing of the gate voltage/current, see Fig. 20.31.

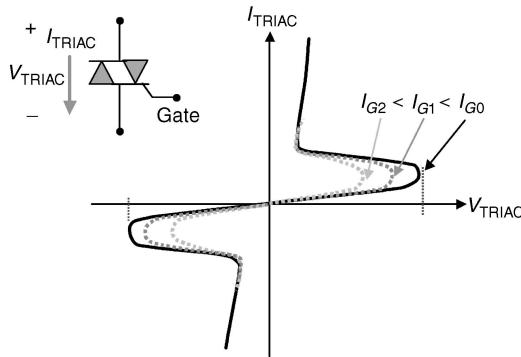



FIGURE 20.32 Triac characteristic.

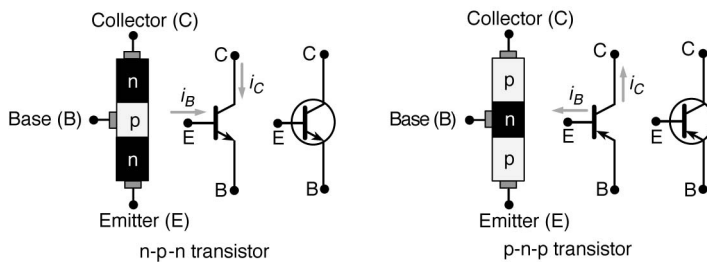



FIGURE 20.33 Biploar junction transistors (BJTs).

Triac

The thyristor can only be turned on in the forward biased direction. A *triac* is a controlled switch that is equivalent to a pair of thyristors that are connected in an anti-parallel configuration, see Fig. 20.32. As depicted in Fig. 20.32, a triac can be turned on in both the reverse and forward directions.

Transistors

A *transistor* is a semiconductor device that has three or more terminals and can provide power amplification and switching. As we have seen in the previous discussions, electronic switching can be accomplished through either diodes or thyristors. Diode switching does not provide any control freedom. A thyristor is a three-terminal device and the third (gate) terminal can be used to control and switching instant. However, one drawback of thyristor switching is that the switching control is only in one direction, i.e., the gate terminal can only be used to turn on the device. The thyristor switch can only be turned off by dropping the anode current to zero.

A transistor is a special semiconductor device that can be used for *power amplification* by modulating a relatively large current between or voltage across two terminals using a small control current or voltage, and *switching* by effectively opening and closing the connection between two terminals using a controlled signal on the third terminal.

Transistors form the basis of modern electronics and are the fundamental building blocks for digital electronics, operational amplifiers, and power electronics. There are three common types of transistors, bipolar junction transistors (BJTs), metal-oxide field effect transistors (MOSFETs), and insulated gate bipolar transistors (IGBTs).

Bipolar Junction Transistors (BJTs)

A bipolar junction transistor is a three-layer device that is made of the p-n-p or the n-p-n combinations of semiconductors, see Fig. 20.33. BJTs have three terminals connected to each of the three layers called

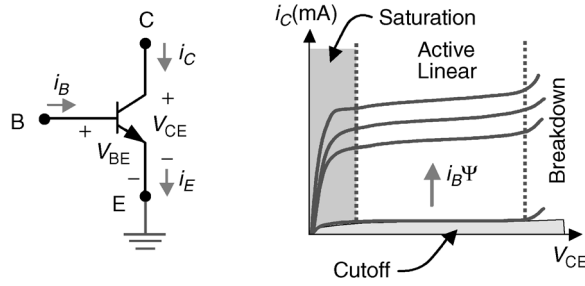



FIGURE 20.34 Characteristics of a common emitter n-p-n BJT.

collector (C), emitter (E), and base (B). Figure 20.34 shows the operation of an n-p-n type BJT under a common emitter type connection. BJTs can operate in three regions:

1. *Cutoff*—When the base-emitter voltage is less than the turn-on voltage V_γ, the base current i_B will be negligible. The transistor is in the *cutoff* region and no current will flow through the collector and emitter terminal, i.e.,

$$\begin{cases} V_{\text{BE}} < V_\gamma \\ i_B = 0 \end{cases} \Rightarrow \begin{cases} i_C \approx 0 \\ V_{\text{CE}} \geq 0 \end{cases}$$

   Typically, $V_\gamma = 0.6 - 0.7$ V. In this mode, the transistor from C to E can be viewed as an open connection. This is analogous to the closed flow control valve.

2. *Active Linear*—When $V_{\text{BE}} = V_\gamma$, the transistor is in the *active linear* region, where

$$V_{\text{BE}} = V_\gamma \quad \text{and} \quad \begin{cases} i_C = \beta \cdot i_B \\ V_{\text{CE}} > V_\gamma \end{cases}$$

   In this mode, the transistor can be viewed as a current-controlled current amplifier, where the collector current i_C is proportional to the base current i_B. The proportionality constant (current amplification factor or current gain) $\beta = 20 \sim 200$, is often denoted as h, hf, or h_{FE} in the data sheets. In this mode, the connection between the terminals C and E can be viewed as closed. This is analogous to a partially opened flow control valve, where the amount of the fluid (current) flow is proportional to the size of the valve opening (base current magnitude). The power dissipation across the transistor P_{BJT} is

$$P_{\text{BJT}} = i_C \cdot V_{\text{CE}}$$

3. *Saturation*—When the base current i_B is larger than the maximum available collector current i_C, the transistor is in the *saturation* region, where

$$\begin{cases} i_B > i_C/\beta \\ V_{\text{BE}} = V_\gamma \end{cases} \quad \text{and} \quad V_{\text{CE}} = V_{\text{SAT}} \approx 0.2\ V \tag{20.13}$$

   In this mode, the transistor can be viewed as a closed switch between the terminals C and E. The collector current i_C is controlled (determined) by the collector circuit. This is analogous to a completely opened flow control valve, where the flow is determined by the source and the load. Note also that when the transistor is in saturation, the collector-emitter voltage drop is maintained at a small value called the saturation voltage V_{SAT}.

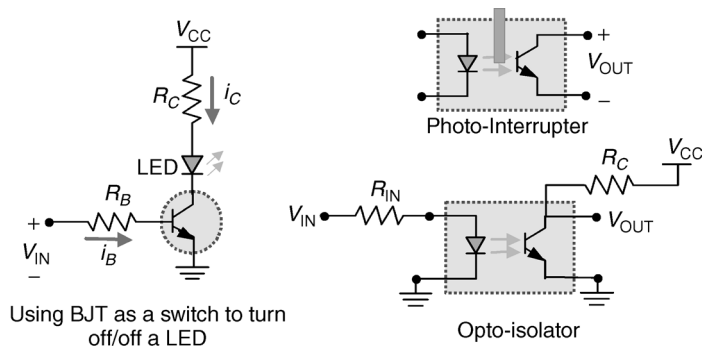



FIGURE 20.35 Some examples of using BJT.

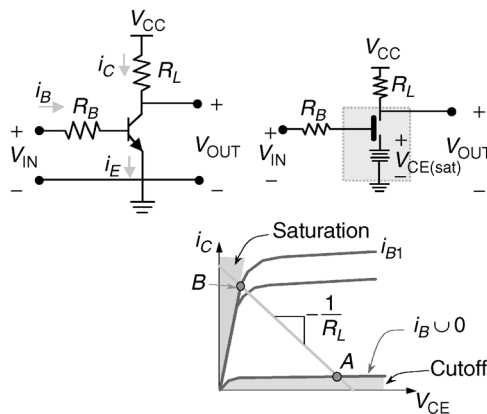



FIGURE 20.36 BJT as a current controlled switch.

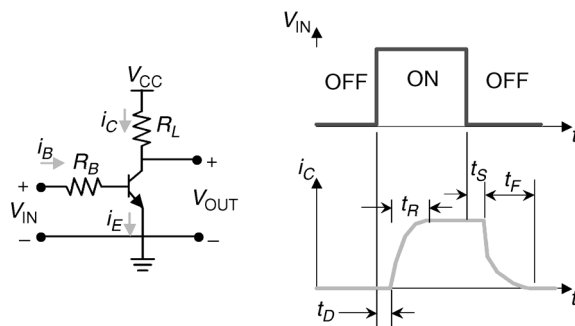



FIGURE 20.37 BJT switching characteristics.

In summary, when the transistor is saturated, it acts as a closed switch. When a transistor is in the cutoff region, it acts as an open switch. When it is in the active region, it acts as a current (i_B) controlled current (i_C) amplifier. Although this is a very simplistic approximation, it is very useful for designing and understanding power electronics and interfacing electromechanical systems. Figure 20.35 illustrates some examples of BJT devices and applications.

When carefully controlling the base-emitter voltage V_{BE} and base current i_B, the transistor can be made to operate between the cutoff and the saturation region, which act as a switch, see Fig. 20.38. Realistically, transistor switching is not instantaneous (see Fig. 20.37). The turn-on time t_{ON} of the transistor is the

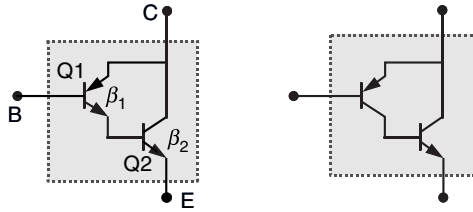



FIGURE 20.38 Two type of Darlington transistor pairs.

sum of the delay time t_D and the rise time t_R. Similarly, the turn-off time t_{OFF} is the sum of the storage time t_S and the fall time t_F. The turn-on and turn-off time of a transistor limits the maximum switching frequency. Typical switching frequency for a power BJT is between 2 and 20 kHz. Generally speaking, BJTs can switch at a higher frequency than thyristors but can handle less power. Power BJTs can handle currents up to several hundred amperes and V_{CE} up to about 1 kV.

Power dissipation is a key design constraint for BJTs. Recall that if the BJT is used in the active linear region (linear amplifier), the power dissipation is $P_{BJT} = i_C \cdot V_{CE}$ with $V_{CE} > V_\gamma$. With a large collector current and considering the small volume and thermal mass of the device, the transistor is not very efficient when operating in the active linear region. On the other hand, when the BJT is switching between saturation and cutoff, the collector current will be small (during cutoff) and V_{CE} will be small (during saturation). The switching power dissipation is much smaller compared with the active linear mode of operation. This makes switching much more efficient.

One design consideration working with BJT is to supply adequate base current, especially when the transistor is to operate in the saturated region, see Eq. (20.13). This may require large input power and may overload the input stage. As will be discussed later, this is also the main reason that BJTs are less used in switching power electronics and are being replaced by devices such as MOSFET and IGBT, which require much less control current. One solution to this constraint is to increase the current gain β. A simple and elegant implementation to increase the effective current gain of a BJT is the Darlington pair configuration.

Darlington Transistor Pairs

A *Darlington transistor pair* connects two BJT transistors to form an effective three terminal device that has increased current gain, see Fig. 20.38. In Fig. 20.38, let β_1 and β_2 be the current gains of the two transistors, then the relationship between the base current of transistor Q1 and the collector current of transistor Q2 is

$$i_{C2} = \beta_2 \cdot i_{B2} = \beta_2 \cdot (\beta_1 \cdot i_{B1}) = (\beta_2 \cdot \beta_1) \cdot i_{B1} = \beta_D \cdot i_{B1}$$

Therefore, the effective current gain for the Darlington transistor pair is the product of the two individual current gains, i.e., $\beta_D = \beta_1 \cdot \beta_2$. For a typical Darlington pair, this can be in the range of 500–10,000. The trade-off for using Darlington pair configuration is the additional space (real estate) needed for two transistors instead of one.

Metal-Oxide-Semiconductor Field Effect Transistor (MOSFET)

MOSFET is a type of field effect transistor (FET). FETs are voltage controlled three terminal devices respectively called drain (D), source (S), and gate (G). The terms come from the analogy of overhead tank system that uses a gate valve to control the water flow from source to drain. MOSFET uses a metal plate as the gate terminal and it is insulated from the p- or n-type silicon substrate by a thin layer of oxide (see Fig. 20.39). When a gate voltage V_G is applied to the gate plate, an electrostatic field induces reverse charges at the gate and the substrate. The charges at the substrate initiate transistor type characteristics by forming either an n-type channel or a p-type channel. Hence, the n- or p-type MOSFET classifications (see Fig. 20.39).

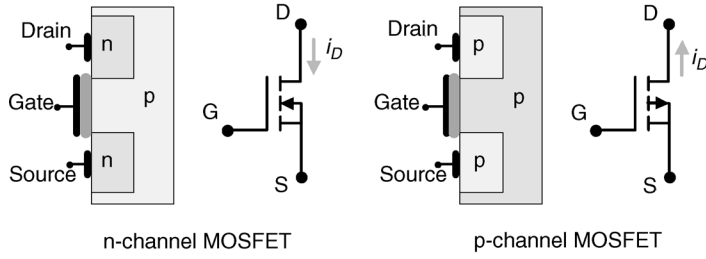



FIGURE 20.39 Metal-oxide-semiconductor (MOS) field effect transistor (FET).

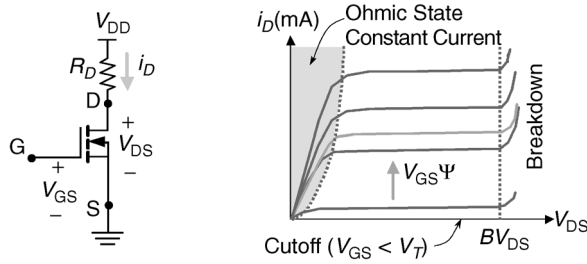



FIGURE 20.40 Enhancement mode MOSFET characteristic.

For a majority of the power amplification and modulation applications, MOSFETs are designed to operate in the *enhancement* mode. Figure 20.42 illustrates the *enhancement* mode characteristic of an n-channel MOSFET:

1. *Cutoff*—When the potential across the gate and the substrate (source) V_{GS} is less that the turn-on (threshold) voltage V_T, the MOSFET is in the *cutoff* region and there is negligible current flow through the drain (D) terminal, i.e.,

$$\begin{cases} V_{GS} < V_T \\ i_G = 0 \end{cases} \Rightarrow \begin{cases} i_D \approx 0 \\ V_{DS} \approx V_{DD} \end{cases}$$

   Typically, $V_T \approx 1–2$ V. In this mode, the transistor from D to S can be viewed as an open connection.

2. *Active Region*—When the $V_{GS} > V_T$, the MOSFET is in the *active* region, where

$$V_{GS} > V_T \quad \text{and} \quad \begin{cases} i_D \propto (V_{GS} - V_T)^2 \\ V_{DS} > V_{GS} - V_T \end{cases}$$

   In this mode, the transistor can be viewed as a voltage-controlled current amplifier, where the drain current i_C is proportional to square of the difference between the gate-source voltage and the threshold voltage. The drain current is controlled by the gate-source voltage V_{GS}. The power dissipation across the transistor P_{FET} is

$$P_{FET} = i_D \cdot V_{DS}$$

3. *Ohmic State*—When V_{GS} is large enough so that the drain current is determined by the drain source circuit, the MOSFET is in *saturation* and

$$V_{GS} \gg V_T \quad \text{and} \quad \begin{cases} i_D = V_{DD}/R_D \\ V_{DS} \approx i_D \cdot R_{ON}(V_{DS}) < V_{GS} - V_T \end{cases} \tag{20.14}$$

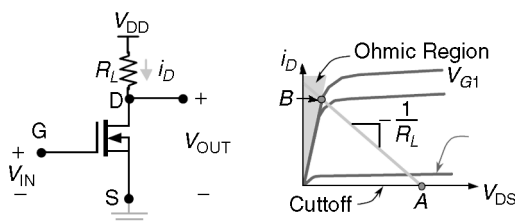



FIGURE 20.41 MOSFET as a voltage controlled switch.

In this mode, the transistor can be viewed as a closed switch between the terminals D and S with a voltage controlled resistance R_{ON}. The drain current i_D is controlled (determined) by the drain circuit. At rate current, the V_{DS} drop during saturation ranges from 2 to 5 V.

When operating in the enhancement mode, a MOSFET behaves very similar to a BJT. Instead of base current, the MOSFET behavior is determined by the gate voltage. When carefully controlling the gate voltage of a MOSFET, the transistor can be made to operate as a voltage controlled switch (Fig. 20.41) that operates between the cutoff (point A) and the Ohmic (point B) region.

One advantage of a MOSFET device is that the MOSFET has significantly larger input impedance as compared to BJT. This simplifies the circuit that is needed to drive the MOSFET since the magnitude of the gate current is not a factor. This also implies that a MOSFET is much more efficient than BJTs as well as it can be switching at a much higher frequency. Typical MOSFET switching frequency is between 20 and 200 kHz, which is an order of magnitude higher than BJTs. Power MOSFETs can carry drain currents up to several hundreds of amperes and V_{DS} up to around 500 V.

Field effect is one of the key reasons why MOSFET has better switching performance than BJT. However, static field is also one of its main failure modes. MOSFETs are very sensitive to static voltage. Since the oxide insulating the gate and the substrate is only a thin film (in the order of a fraction to a few micrometer), high static voltage can easily break down the oxide insulation. A typical gate breakdown voltage is about 50 V. Therefore, static electricity control or insulation is very important when handling MOSFET devices.

Comparing BJT with MOSFET, we can conclude the following:

- Both can be used as current amplifiers.
  - BJT is a current-controlled amplifier where the collector current i_C is proportional to the base current i_B.
  - MOSFET is a voltage-controlled amplifier where the drain current i_D is proportional to the square of the gate voltage V_G.
- Both can be used as three terminal switches or voltage inverters.
  - BJT: switching circuit give rise to TTL logics.
  - MOSFET: switching circuit give rise to CMOS logics.
- BJT usually has larger current capacity than similar sized MOSFET.
- MOSFET has much higher input impedance than BJT and is normally off, which translates to less operating power.
- MOSFETs are more easily fabricated into integrated circuit.
- MOSFETs are less prone to go into thermal runaway.
- MOSFETs are susceptible to static voltage (exceed gate breakdown voltage ~50 V).
- BJT has been replaced by MOSFET in low-voltage (<500 V) applications and is being replaced by IGBT in applications at voltages above 500 V.

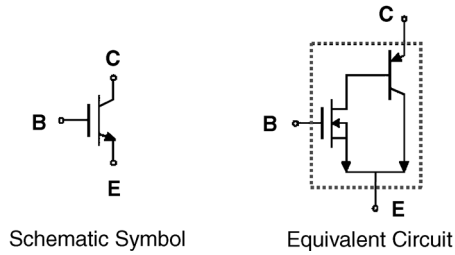



FIGURE 20.42 Insulated gate bipolar transistor (IGBT).

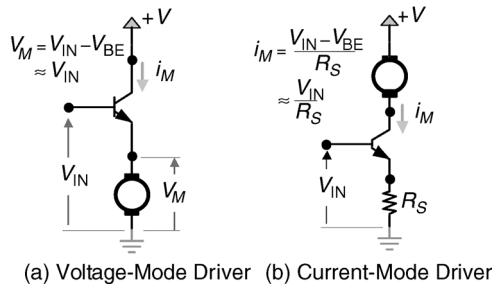



FIGURE 20.43 Two basic linear motor amplifiers.

Insulated Gate Bipolar Transistor (IGBT)

IGBT is a voltage-controlled transistor that has the terminals identified in the same way as BJTs. IGBT is a four-layer device that has the similar construction of a MOSFET with an additional *p* layer. Figure 20.42 shows the schematic symbol and equivalent circuit for an IGBT. IGBT has the combined characteristics of the BJT and MOSFET. Similar to MOSFET, it has high input impedance and high switching frequency. It also has high power handling capacity like the BJT.

Typical Power Amplifiers for Electromechanical Actuators

Power amplification and modulation for electromechanical actuators are classified into two basic categories, based on the methods the respective power electronics are driven. Linear amplifiers drive the BJTs in their active linear region. Switching amplifiers drive the transistors in on-off switch mode. Depending on the control objective, the command signal (Fig. 20.2) to the amplifier can be either a voltage or current command that intends to modulate the electric energy delivered to the energy conversion device. Since most of the electromechanical actuator involves driving an inductance load such as a coil winding of an electromagnet or the rotor of a DC motor, in the following discussion, we will use the DC motor as an example of an inductance load for the power amplifier.

Linear Amplifiers

Figure 20.43 shows the basic drive circuit for linear voltage and current control amplifiers. Both schemes have the following commonalities:

1. The input command voltage $V_i(t)$ is applied to the base of the transistor.
2. The electric power needed to driver the load is provided by a DC supply.
3. The transistors are driven in the active linear region.

Voltage Control (Mode) Amplifier

In Fig. 20.43(a), the motor is driven as a load of an emitter circuit. If the base-emitter voltage is ignored, the voltage across the motor V_M is directly controlled by the input voltage V_{IN} and the current is supplied by the power supply. The amount of current, on the other hand, depends on the applied voltage, speed, and the motor parameter.

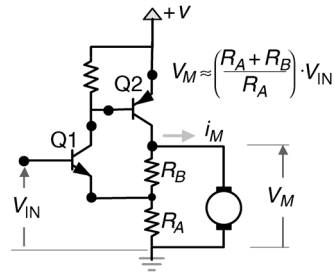



FIGURE 20.44 Variable gain voltage-mode amplifier.

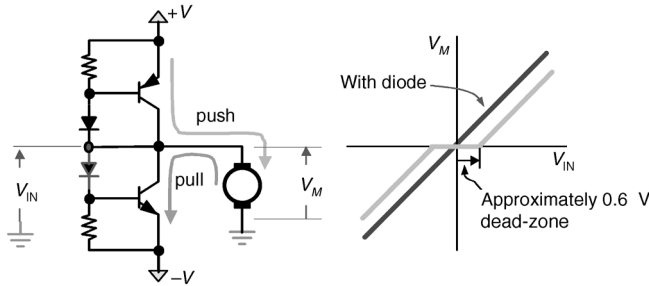



FIGURE 20.45 Bipolar voltage-mode amplifier.

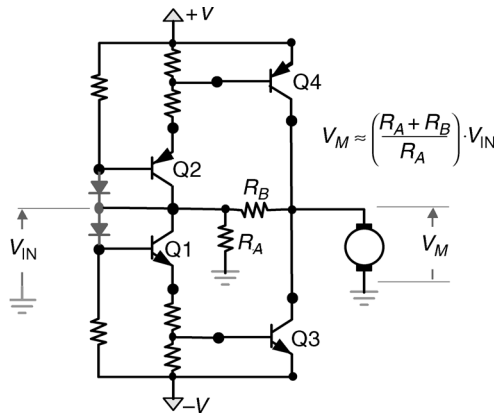



FIGURE 20.46 Bipolar variable gain voltage-mode amplifier.

To provide adjustable voltage gain, Fig. 20.44 shows a variable gain voltage-mode amplifier, where the apparent amplifier gain from the command input V_{IN} to the applied motor (winding) voltage V_M can be approximated by $(R_A + R_B)/R_A$, which can be adjusted by proper selection of the resistors R_A and R_B. If a large motor current is required, transistor Q2 can be replaced by a Darlington transistor pair.

The amplifiers shown in Fig. 20.43 and Fig. 20.44 can only drive the current through the motor (load) in one direction. Hence, they are also called *unipolar* amplifiers. To provide bidirectional current flow, two transistors can be connected with the motor in a *push-pull* type configuration, as shown in Fig. 20.45. The two diodes in the circuit are used to eliminate the dead-zone created by the base-emitter voltage drop for the transistors. Notice that a bipolar voltage source is needed for this configuration. Figure 20.46 shows a bipolar voltage-mode driver with variable gain $(R_A + R_B)/R_A$. Similarly, if larger motor current is required, transistors Q3 and Q4 can be replaced by Darlington transistor pairs.

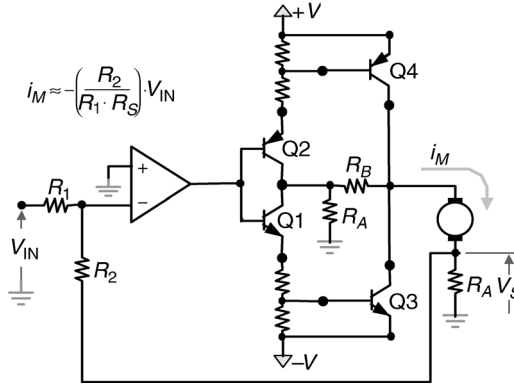



FIGURE 20.47 Bipolar variable gain current-mode amplifier.

Current Control (Mode) Amplifier

As previously discussed, in many electromagnetic actuators, the output force or torque of the device has strong correlation with the winding current, e.g., for a permanent magnet DC motor and a voice coil actuator, the output torque and force are proportional to the input current. Therefore, in many motion control applications, it is more desirable to have a voltage-to-current conversion (*current-mode amplifier*) at the power stage, where the input voltage command is proportional to the current flowing into/out of the motor (winding). Figure 20.43(b) shows a basic circuit for a current-mode amplifier. The relationship between the emitter (motor) current i_M and the input voltage command V_{IN} is

$$i_M = \frac{V_{IN} - V_{BE}}{R_S}$$

If the base-emitter voltage is ignored, the voltage across the motor current i_M is proportional to the input voltage V_{IN}, i.e., $i_M \approx (1/R_S) \cdot V_{IN}$.

Figure 20.47 shows a basic bipolar current-mode amplifier. An Op-Amp is used to close the current loop. The resistor R_S, often called the *sensing resistor*, is used to sense the motor current for feedback to the Op-Amp. Depending on the desired current magnitude, the sensing resistor needs to have adequate power rating to dissipate the heat ($i_M^2 \cdot R_S$) generated by flowing current through the resistor. For a zeroth order approximation, at steady state, the Op-Amp will try to equalize the potential at the positive and the negative terminals, i.e., it will try to make

$$V_S \approx -\left(\frac{R_2}{R_1}\right) \cdot V_{IN} \qquad ,$$

which implies

$$i_M \approx -\left(\frac{R_2}{R_S \cdot R_1}\right) \cdot V_{IN} \qquad .$$

Although a current amplifier tends to have a linear relationship between the command input and the winding current, there is practical limitation due to the limited source voltage. In Fig. 20.47, the supply voltage is ±V. Assuming that the motor winding has resistance R_M, the maximum current i_{MAX} the voltage source can supply is upper bounded by

$$i_{MAX} < \frac{V}{R_M + R_S} \qquad .$$

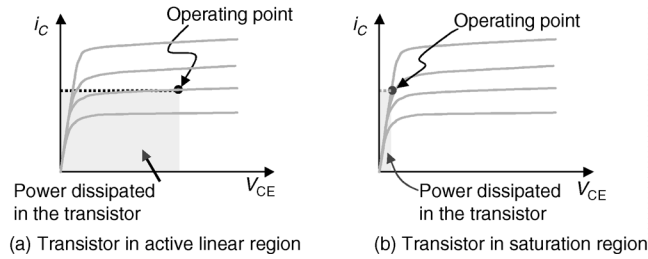



FIGURE 20.48 Power dissipation in transistors.

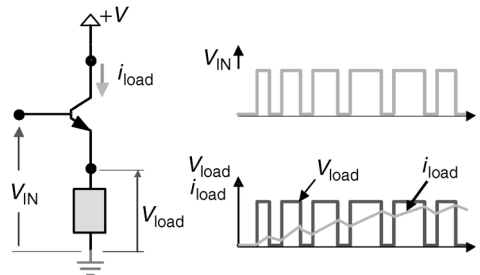



FIGURE 20.49 Simple switching amplifier with switching input.

The above bound has not considered the effect of the back-emf that will be induced in the winding if the winding is moving. Hence, the amount of current available for a current-mode amplifier is limited and needs to be considered when working with a current-mode amplifier.

Switching Amplifiers

Linear amplifiers are simple and do not generate electrical noises. However, since the final stage transistors are operating in the active linear region, significant power is dissipated into heat; this reduces the efficiency of the device as well as requires large heat sinks to protect the components. However, as shown in Fig. 20.48, when operating in the saturation region, if the collector-emitter voltage drop is in the order of 1 V or less, the power loss across the transistor is significantly less, given the same amount of current flow. The trade-off is that additional circuits are needed to provide the modulation for current or voltage control.

Figure 20.49 shows a simple switching amplifier that is simply a transistor connecting a load. It is essentially the same as the basic linear amplifier shown in Fig. 20.43(a). The difference is in the way the transistor is controlled. For a switching amplifier, the input (base) voltage only takes on two values (states), high and low. When the base (input) voltage is high, the transistor is turned on in the saturation mode and current will flow through the load. If we neglect the collector-emitter voltage drop, the voltage across the load is approximately the supply voltage. When the base voltage is low, the transistor is turned off in the cutoff state and no voltage is applied to the load. If the load has a low pass characteristic, the average current/voltage across the load will be proportional to the turn-on time. Therefore, if the switching frequency is sufficiently high (relative to the load impedance), the effective voltage/current across the load can be modulated by the percent high input voltage, e.g., if the V_{IN} is high 80% of the time, the average voltage across the load will be close to 80% of the supplied voltage V. This is the so-called *pulse-width modulation* (PWM). Another benefit of using switching amplifiers is that V_{IN} can be directly interfaced with a digital device without the need for a DAC.

Push-Pull (Class B) Power Amplifier

The switching amplifier shown in Fig. 20.49 is unipolar, i.e., it can only drive current through the load in one direction. Figure 20.52 shows a simple *push-pull* (*Class B*) type power stage to supply bi-directional current to the load. The circuit in Fig. 20.50 is very similar to the bipolar voltage-mode amplifier shown

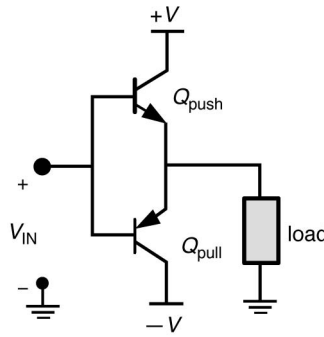



FIGURE 20.50 Switching push-pull amplifier.

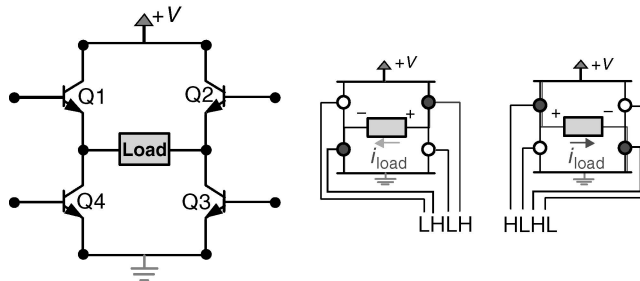



FIGURE 20.51 H-bridge driver.

in Fig. 20.45. The difference is also in the way the transistors are controlled. When the base voltage V_{IN} is sufficiently positive (+V), the push transistor Q_{push} will be turned on and the pull transistor Q_{pull} will be turned off. This results in a load current flowing from positive supply to ground. If the base voltage is sufficiently negative (–V), Q_{push} will be turned off and Q_{pull} will be turned on, which results in a current flow from ground to the negative supply. To modulate the load voltage/current, PWM can also be used. This configuration is also called a *half H-bridge* driver or half-bridge driver for short. From an implementation perspective, this device requires both a positive supply and a negative supply, which tends to increase the complexity and cost of the circuit.

H-Bridge Driver

H-bridge configuration is a neat solution to achieve bipolar operation with unipolar supply. Figure 20.51 shows a simple H-bridge circuit driving a load. An H-bridge consists of four transistors that are connected in a Wheatstone bridge configuration. By turning on/off different pairs of transistors (Q1-Q3) or (Q2-Q4), bipolar voltage across the load can be achieved using a unipolar supply, see Fig. 20.51. In many applications, the transistors pairs in the H-bridge can be directly driven by the output of a digital device (TTL or CMOS). The n-p-n or n-channel transistors can be turn on to saturation by a high output from the digital port and turned off by a low output. If large amount of current is required for the load, Darlington pairs can be used in place of the individual transistors. Since MOSFETs have larger input impedance and faster switching characteristics, they are replacing BJTs in almost all switching applications.

Pulse-Width Modulation (PWM)

PWM is one of the more common ways of encoding analog information using digital signal. A PWM signal is a wave of fixed frequency and varying duty cycle (pulse width). The duty cycle in PWM context refers to the percentage of time that the signal is in the active state—usually this means a state of logic 1, see Fig. 20.52. In essence, PWM encodes (modulates) the information in the time domain rather than the voltage domain as with analog signals.

PWM actuation has several advantages over the use of D/A converters and linear components. One is the efficiency where the switching amplifiers are more efficient than their linear counter parts. Another

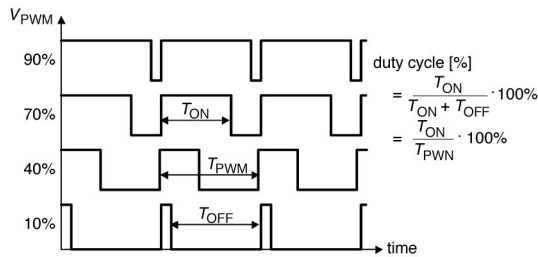



FIGURE 20.52 Pulse-width modulation signals.

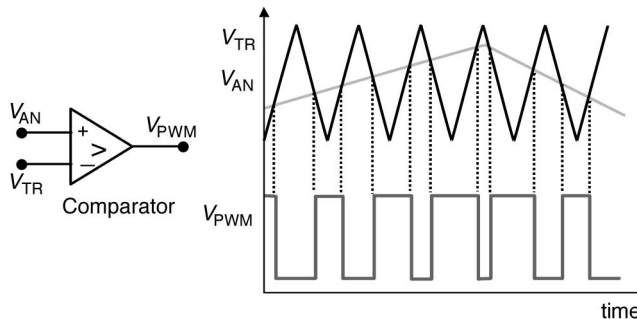



FIGURE 20.53 Generating PWM signal from analog signal.

advantage is there is no need for D/A conversion. Digital signal is maintained from the microprocessor/microcontroller to the power amplifier. In additional to having better noise rejection capability, this also reduces the need for a DAC and tends to make the circuit simpler and more cost effective. One drawback for using PWM and switching amplifiers as a whole is that the high frequency switching induces radio frequency interference (RFI) and electromagnetic interference (EMI).

The fixed PWM carrier frequency is one main design consideration. Ideally, the PWM frequency should be high enough to avoid generating audible switching noise, which mean that it should be greater than 20 kHz. However, there are a few factors that put an upper bound on the carrier frequency. Switching losses of switching devices tends to increase as the switching frequency increases. This reduces the efficiency of switching components and amplifiers. Higher PWM carrier frequency requires faster switching components that cost more. The amount of current going through the device also limits the switching rate. In general, sub-horsepower devices and office/desktop equipments usually use PWM at 20–40 kHz. For larger scale industrial applications, the PWM frequency tends to be less than 500 Hz. Another commonly specified design parameter is the PWM resolution. This is required for generating PWM from a digital source. The PWM resolution is equivalent to the quantization resolution for ADC. An 8-bit PWM means that there are $2^8 = 256$ different pulse widths per PWM carrier signal period.

PWMs are widely adopted in the field and almost all microcontrollers and microprocessors have at least one PWM output port. PWM signal can be easily generated from analog signal by comparing the analog signal with a periodic triangular signal through a comparator, see Fig. 20.53.

Interfacing Considerations

We will conclude this section by discussing some issues relating to interfacing between the electromechanical actuator and the power amplification device.

Driving Inductive Load

A majority of the electromechanical actuators use coils (windings) to convert electrical energy to magnetic energy. From a power driver viewpoint, windings are resistive and inductive loads. Inductors are energy

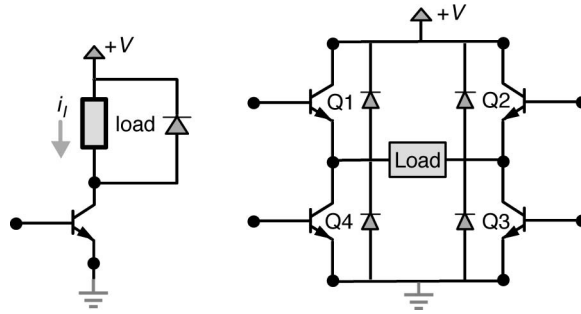



FIGURE 20.54 Using diodes to reduce swithcing voltage when driving inductive loads.

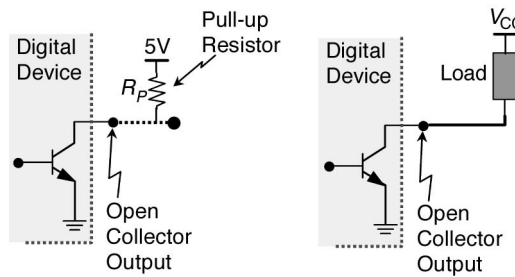



FIGURE 20.55 Open-collector output.

storage elements, where the energy is stored in the induced magnetic field. The voltage across an ideal inductor $V_I(t)$ is

$$V_I(t) = L \cdot \frac{d}{dt} i_I(t) \tag{20.15}$$

where $i_I(t)$ is the current going through the inductor and L is the inductance. When the current to the inductor is suddenly switched off, e.g., by switching off a driving transistor, Equation (20.15) indicates that there will be a large transient voltage build-up across the inductor. If not properly suppressed this transient voltage can shorten or even damage the driving transistor. This is sometimes called *inductor kickback*.

A simple method of reducing the instantaneous switching voltage surge is to create a loop for the excess energy to flow. This can be done by placing diodes in parallel with the load, see Fig. 20.54. Figure 20.54 illustrates two methods of using *flyback* or *free-wheeling* diodes to suppress switching voltage surge when driving inductive loads.

Open-Collector Output

For some digital devices, the output stage (pin) is simply the collector of a transistor. This is called an *open-collector* output, see Fig. 20.55. Since the output of the device is only the collector of a transistor, it has no output drive capacity. The output value can be measured through a pull-up resistor, see Fig. 20.55. Open-collector output is convenient for driving electromechanical devices if the output transistor can sink adequate current, see Fig. 20.57.

Isolation

Recall that the power amplification/modulation part of an electromechanical actuator contains both low- and high-energy signals, see Fig. 20.2. For safety and reliability reasons, it is desired to prevent transients or noise spikes in the high power side of the system from the signal processing (low power) side of the circuit. Mechanical relay is one option. *Optoisolators* or *optocouplers* use light to couple the high and low

energy side of the device. Typically, an LED source is combined with either a phototransistor or photo thyristor, see Fig. 20.35. In addition to signal isolation, optoisolators also help to reduce ground loop issues between the logic and power side of the circuit.

Grounding

It is important to provide common ground among the different devices. For electromechanical actuators, the high energy side is often switching at high frequency; if the ground point of the high energy side of the circuit is directly connected to the ground of the low energy side of the circuit, switching noise may propagate through the ground wire and negatively affect the operation of the low energy side of the system. It is recommended that separate common grounds are established for the high and low energy side and the two grounds are then connected at the power supply. In addition, an adequate-sized ground plane needs to be provided to minimize the possibility of differences among grounding points.

20.2 Electrical Machines

C. J. Fraser

The utilization of electric motors as the power source in a mechatronic application is substantial. Electric motors, therefore, often feature as the prime mover in a variety of driven systems. It is usually the mechanical features of the application that determines the type of electric motor to be employed. The torque–speed characteristics of the motor and the driven system are therefore very important. It is perhaps then a paradox that while the torque–speed characteristics of the motor are readily available from the supplier, the torque–speed characteristics of the driven system are often quite obscure.

The dc Motor

All conventional electric motors consist of a stationary element and a rotating element, which are separated by an air gap. In dc motors, the stationary element consists of salient "poles," which are constructed of laminated assemblies with coils wound round them to produce a magnetic field. The function of the laminations is to reduce the losses incurred by eddy currents. The rotating element is traditionally called the "armature" and this consists of a series of coils located between slots around the periphery of the armature. The armature is also fabricated in laminations, which are usually keyed onto a location shaft. A very simple form of dc motor is illustrated in Fig. 20.56.

The single coil is located between the opposite poles of a simple magnet. When the coil is aligned in the vertical plane, the conventional flow of electrons is from the positive terminal to the negative terminal. The supply is through the brushes, which make contact with the commutator segments. From Faraday's laws of electromagnetic induction, the "left-hand rule," the upper part of the coil will experience a force acting from right to left. The lower section will be subject to a force in the opposite direction. Since the

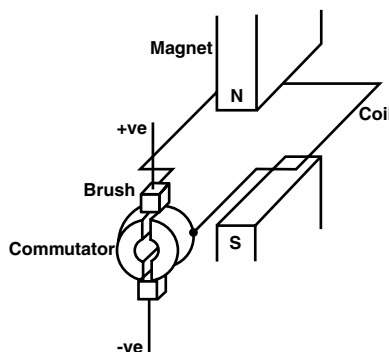



FIGURE 20.56 Single-coil, 2-pole dc motor.

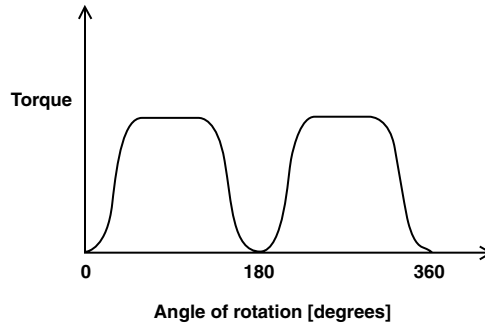



FIGURE 20.57 Torque variation through one revolution.

coil is constrained to rotate, these forces will generate a torque, which will tend to make the coil turn in the anti-clockwise direction. The function of the commutator is to ensure that the flow of electrons is always in the correct direction as each side of the coil passes the respective poles of the magnet. The commutator incorporates brass segments, separated by insulating mica strips. The carbon brushes make sliding contact with the commutator.

When the coil lies in the horizontal direction, there is maximum magnetic flux linking the coil but a minimum rate of change of flux linkages. On the other hand, when the coil is in the vertical plane, there is zero flux linking the coil but the rate of change of flux linkages is a maximum. The resultant change in torque acting on the coil through one revolution is as shown in Fig. 20.57.

If two coils physically displaced by 90° are used in conjunction with two separate magnets, also displaced by 90°, then the output torque is virtually constant. With the introduction of a second coil, the commutator needs to have four separate segments. In a typical dc machine there may be as many as 36 coils, which would require a 72-segment commutator.

The simple dc motor of Fig. 20.56 can be improved in perhaps three obvious ways. Firstly, the number of coils can be increased, the number of turns in each coil can be increased, and finally the number of magnetic poles can be increased. A typical dc machine would therefore normally incorporate four poles, wired in such a way that each consecutive pole has the opposite magnetic polarity to each of its immediate neighboring poles. If the torque generated in the armature coils are to assist one another then while one side of the coil is passing under a north pole, the other side must be passing under a south pole. With a two-pole machine the armature coils are wound with one side of the coil diametrically opposite the other. In a four-pole machine the coils are wound such that one side of the coil is displaced 90° from the other. The size of the machine will generally determine how many coils and the number of turns on each coil which can be accommodated.

Armature Electromotive Force (emf)

If a conductor cuts a magnetic flux, a voltage of 1 V will be induced in the conductor if the flux is cut at the rate of 1 Wb/s. Denoting the flux per pole as Φ and the speed (in rev/s), as N, for a single turn coil and two-pole machine, the emf induced in the coil is given as

$$E_{\text{coil}} = \frac{\text{flux per pole}}{\text{time for half rev}} = \frac{\Phi}{1/2N} = 2N\Phi \tag{20.15}$$

For a machine having Z_s armature conductors connected in series, i.e., $Z_s/2$ turns, and $2p$ magnetic poles, the total induced emf is

$$E = \frac{2N\Phi Z_s 2p}{2} = 2N\Phi Z_s p \tag{20.16}$$

The induced emf or back emf will oppose the applied voltage. Since the emf is directly proportional to the motor speed then on startup, there will be no back emf generated. This will have consequences on the current, which will be drawn by the coils, and some measures will have to be taken to counteract this effect. This topic will be considered later.

Armature Torque

The force on a current carrying conductor is given as

$$F = BLI \tag{20.17}$$

where B is the magnetic flux density under a pole, I is the current flowing in the conductor, and L is the axial length of the conductor.

The torque on one armature conductor is, therefore,

$$T = Fr = B_{\mathrm{av}} LI_a r \tag{20.18}$$

where r is the radius of the armature conductor about the center of rotation, I_a is the current flowing in the armature conductor, L is the axial length of the conductor, and B_{av} is the average flux density under a pole.

Given that $B_{\mathrm{av}} = \Phi/[(2\pi rL)/2p]$, the resultant torque per conductor is

$$T = \frac{\Phi 2pLI_a r}{2\pi rL} = \frac{\Phi pI_a}{\pi} \tag{20.19}$$

For Z_s armature conductors connected in series, the total torque (in Nm) on the armature is given by

$$T = \frac{\Phi pI_a Z_s}{\pi} \tag{20.20}$$

Terminal Voltage

Denoting the terminal voltage by V, in normal running conditions we have a balanced electrical system where:

$$V = E + I_a R_a \tag{20.21}$$

Since the number of poles and number of armature conductors are fixed, then from Eq. (20.16) we have a proportionality relationship between the speed, the induced emf, and the magnetic flux, i.e.,

$$N \propto \frac{E}{\Phi} \tag{20.22}$$

Using Eq. (20.21)

$$N \propto \frac{(V - I_a R_a)}{\Phi} \tag{20.23}$$

Since the value of $I_a R_a$ is normally less than about 5% of the terminal voltage then to a reasonable approximation

$$N \propto \frac{V}{\Phi} \tag{20.24}$$

Similarly Eq. (20.19) provides a proportionality relationship between the torque, the armature current, and the magnetic flux, i.e.,

$$T \propto I_a \Phi \tag{20.25}$$

Equation (20.24) shows that the speed of the motor is directly proportional to the applied voltage and inversely proportional to the magnetic flux. All methods of speed control for dc motors are based on this proportionality relationship.

Equation (20.25) indicates that the torque of a given dc motor is directly proportional to the product of the armature current and the flux per pole. It is obvious therefore that speed control methods which are based on altering the magnetic flux will also have an effect on the output torque.

Methods of Connection

The Shunt-Wound Motor

The shunt-wound motor (Fig. 20.58) is wired such that the armature and field coils are connected in parallel with the supply.

Under normal operating conditions, the field current will be constant. As the armature current increases, the armature reaction effect will weaken the field and the speed will tend to increase. However, the induced voltage will decrease due to the increasing armature voltage drop and this will tend to decrease the speed. The two effects are not self cancelling and overall the motor speed will fall slightly as the armature current increases.

The motor torque increases approximately linearly with the armature current until the armature reaction starts to weaken the field. These general characteristics are shown in Fig. 20.59 where it can also be seen that no torque is developed until the armature current is large enough to overcome the constant losses in the machine. Figure 20.60 shows the derived torque-speed characteristic.

Since the torque increases dramatically for a slight decrease in speed, the shunt-wound motor is particularly suitable for driving equipment like pumps, compressors, and machine tool elements where the speed must remain "constant" over a wide range of load conditions.

The Series-Wound Motor

The series-wound motor is shown in Fig. 20.61. As the load current increases, the induced voltage, E, will decrease due to the armature and field resistance drops. Because the field winding is connected in series with the armature, the flux is directly proportional to the armature current. Equation (20.24) therefore

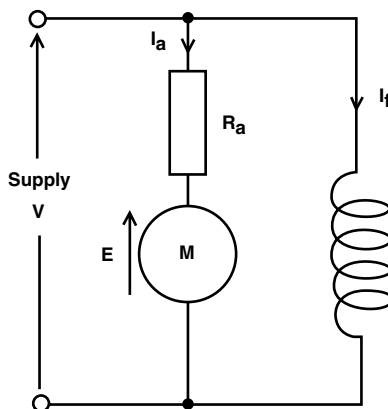



FIGURE 20.58 The shunt-wound motor.

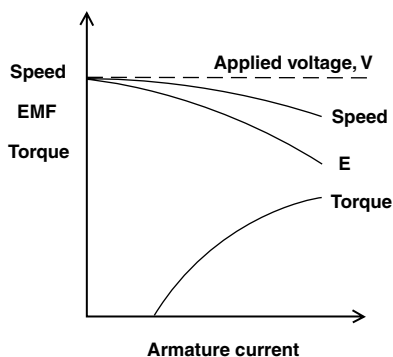



FIGURE 20.59 The shunt-wound motor load characteristics.

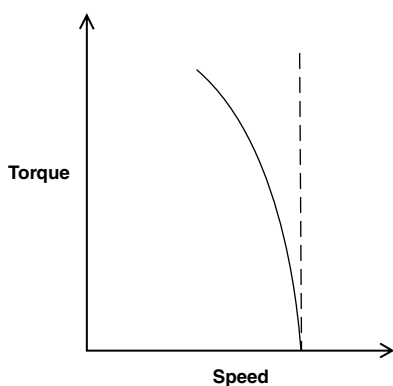



FIGURE 20.60 The shunt-wound torque–speed characteristics.

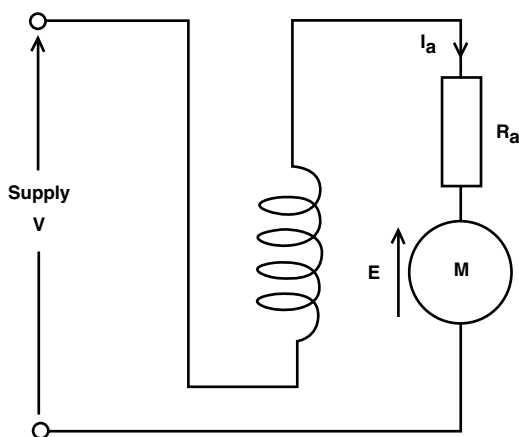



FIGURE 20.61 The series-wound motor.

suggests that the speed–armature current characteristic will take the form of a hyperbola. Similarly, Eq. (20.25) indicates that the torque–armature current characteristic will be approximately parabolic. These general characteristics are illustrated in Fig. 20.62, along with the derived torque–speed characteristic in Fig. 20.63. The general characteristics indicate that if the load falls to a particularly low value

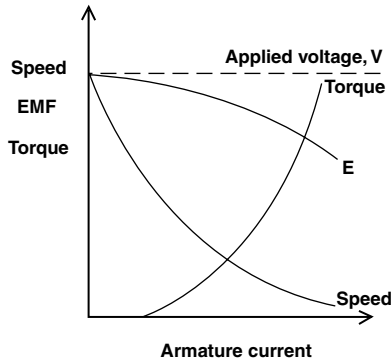



FIGURE 20.62 The series-wound motor load characteristics.

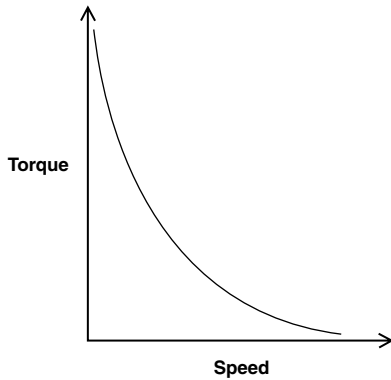



FIGURE 20.63 The series-wound motor torque–speed characteristics.

then the speed may become dangerously high. A series-wound motor should never be used, therefore, in situations where the load is likely to be suddenly relaxed.

The main advantage of the series-wound motor is that it provides a large torque at low speeds. Series-wound motors are eminently suitable, therefore, for applications where a large starting torque is required. This includes, for example, lifts, hoists, cranes, and electric trains.

The Compound-Wound Motor

Compound-wound motors are produced by including both series and shunt fields. The resulting characteristics of the compound-wound motor fall somewhere in between those of the series-wound and the shunt-wound machines.

Starting dc Motors

With the armature stationary, the induced emf is zero. If while at rest, the full voltage is applied across the armature winding then the current drawn would be massive. A typical 40-kW motor might have an armature resistance of about 0.06 Ω. If the applied voltage is 240 V, the current drawn is 4000 A. This current would undoubtedly blow the fuses and thereby cut off the supply to the machine. To limit the starting current a variable external resistance is connected in series with the armature. On start-up the full resistance is connected in series. As the machine builds up speed and increases the back emf, the external resistance can be reduced until at rated speed the series resistance is disconnected. Alternatively, a series resistance can be momentarily activated in conjunction with the starter switch.

Speed Control of dc Motors

Equation (20.24) shows that the speed of a dc motor is influenced both by the applied voltage and the magnetic flux. A change in either one of these parameters will therefore effect a change in the motor speed.

Field Regulator

For shunt-wound and compound-wound motors a variable resistor, called a "field regulator," can be incorporated in series with the field winding to reduce the flux. For the series-wound motor the variable resistor is connected in parallel with the field winding and is called a "diverter." Figures 20.64–20.66 show the various methods of weakening the field flux for shunt-, compound-, and series-wound motors.

In all of the above methods, the flux can only be reduced and from Eq. (20.24) this implies that the speed can only be increased above the rated speed. The speed may in fact be increased to about three or four times the rated speed. The increased speed, however, is at the expense of reduced torque since the torque is directly proportional to the flux which is being reduced.

Variable Armature Voltage

Alternatively, the speed can be increased from standstill to rated speed by varying the armature voltage from zero to rated value. Figure 20.67 illustrates one method of achieving this.

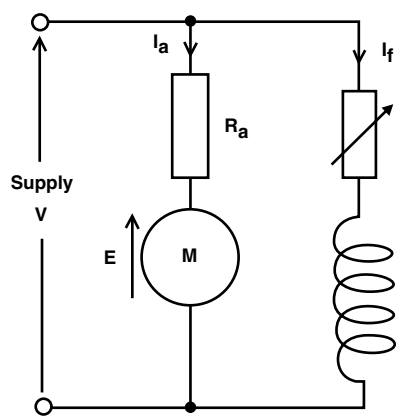



FIGURE 20.64 Speed control by flux reduction: shunt-wound motor.

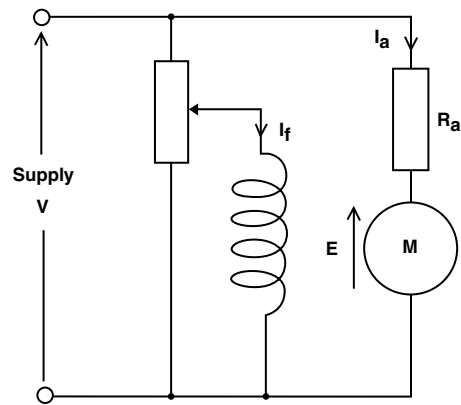



FIGURE 20.65 Speed control by flux reduction: compound-wound motor.

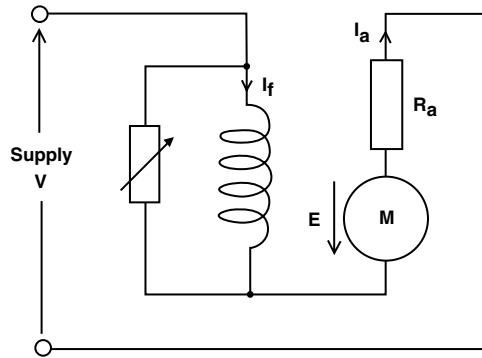



FIGURE 20.66 Speed control by flux reduction: series-wound motor.

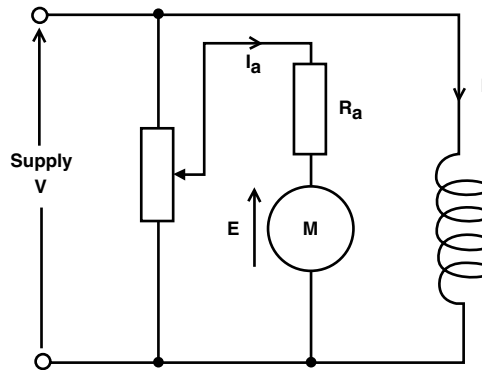



FIGURE 20.67 Speed control by varying the armature voltage.

The potential divider carries the same current as the motor, and this limits this method of speed control to small machines. Additionally much of the input energy is dissipated in the variable resistance, which consequently renders the system inefficient.

Ward Leonard Drive

In this case the variable dc voltage for the speed controlled motor is obtained from a separate dc generator, which is in itself driven by an induction motor.

The field coil for the dc generator is supplied from a center-zero potential divider. When the wiper arm is in the center position, the speed controlled motor is at a standstill. By moving the wiper arm away from the center position the speed of the motor is increased in either clockwise or anti-clockwise direction. The Ward Leonard drive is smooth and accurate in either direction and also provides for very responsive braking. The complexity, however, makes it a very expensive system, and it is only used in high quality applications.

Chopper Control

Figure 20.68 shows a thyristor circuit connected in series with the armature of a dc motor. The thyristor circuit is triggered such that it operates essentially as a high speed ON/OFF switch. The output waveform across the armature terminals is depicted in Fig. 20.69. The ratio of time on to time off, i.e., the "mark/space ratio," can be varied with the result that the average voltage supplied to the armature is effectively varied between zero and fully on. The frequency of the signal may be up to about 3 kHz and the timing circuit is quite complex. Speed control of dc motors using thyristors, however, is effective and relatively inexpensive.

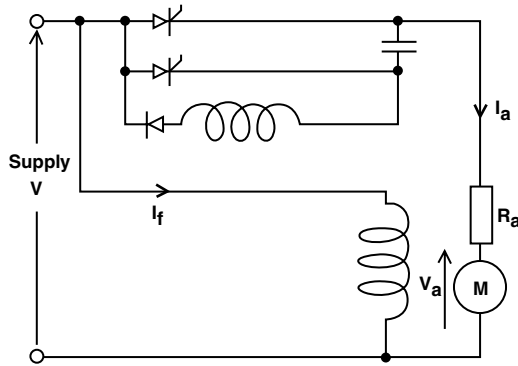



FIGURE 20.68 Speed control using thyristors.

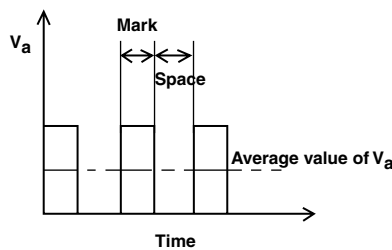



FIGURE 20.69 Voltage across armature terminals.

Efficiency of dc Machines

The losses in dc machines can be generally classified as

1. **Armature losses**: This is the I^2R loss in the armature winding, often referred to as the "copper loss."
2. **Iron loss:** This loss is attributable to magnetic hysteresis and eddy currents in the armature and field cores.
3. **Commutator losses**: This loss is related to the contact resistance between the commutator brushes and segments. The total commutator loss is due to both mechanical friction and a voltage loss across the brushes.
4. **Excitation loss**: In shunt-wound machines, this power loss is due to the product of the shunt current and the terminal voltage.
5. **Bearing friction and windage**: Bearing friction is approximately proportional to the speed, but windage loss varies with the cube of the speed. Both of these losses are fairly minor unless the machine is fitted with a cooling fan, in which case the windage loss can be quite significant.

Despite the variety and nature of the losses associated with dc machines, they have nonetheless a very good performance with overall efficiencies, often in excess of 90%.

AC Machines

Synchronous Motors

Synchronous motors are so called because they operate at only one speed, i.e., the speed of a rotating magnetic field. The production of the rotating magnetic field may be actioned using three, 120° displaced, stator coils supplied with a three-phase current. The rotational speed of the field is related to the frequency

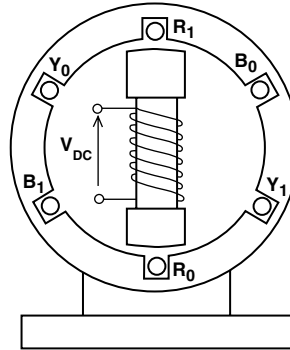



FIGURE 20.70 Simple synchronous motor.

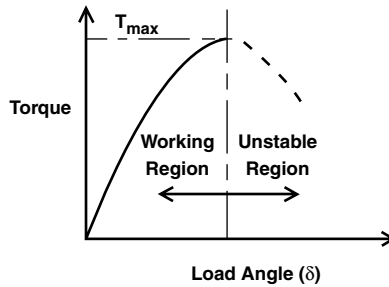



FIGURE 20.71 Torque characteristic for a synchronous motor.

of the currents.

$$N_s = \frac{60f}{\text{number of pole pairs}} \tag{20.26}$$

where N_s is the speed of the field in revolutions per minute and f is the frequency of the supply currents.

The mechanical construction is shown in Fig. 20.70. The rotor field is supplied from a dc source and the stator coils are supplied with a three-phase current. The rotating magnetic field is induced by the stator coils and the rotor, which may be likened to a permanent bar magnet, aligns itself to the rotating flux produced in the stator. When a mechanical load is driven by the shaft, the field produced by the rotor is pulled out of alignment with that produced by the stator. The angle of misalignment is called the "load angle." The characteristics of synchronous motors are normally presented in terms of torque against load angle, as shown in Fig. 20.71.

The torque characteristic is basically sinusoidal with

$$T = T_{\max} \sin \delta \tag{20.27}$$

where $T_{\max}$ is the maximum rated torque and δ is the load angle.

It is evident from Eq. (20.27) that synchronous motors have no starting torque and the rotor must be run up to synchronous speed by some alternative means. One method utilizes a series of short-circuited copper bars inserted through the outer extremities of the salient poles. The rotating magnetic flux induces currents in these "grids" and the machine accelerates as if it were a cage-type induction motor, see following section. A second method uses a wound rotor similar to a slip-ring induction motor. The machine is run up to speed as an induction motor and is then pulled into synchronism to operate as a synchronous motor.

The advantages of the synchronous motor are the ease with which the power factor can be controlled and the constant rotational speed of the machine, irrespective of the applied load. Synchronous motors,

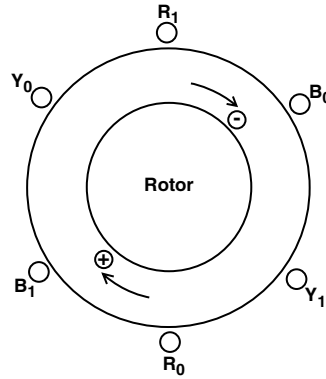



FIGURE 20.72 Schematic representation of an induction motor.

however, are generally more expensive and a dc supply is a necessary feature of the rotor excitation. These disadvantages coupled with the requirement for an independent starting mode make synchronous motors much less common than induction motors.

Induction Motors

The stator of an induction motor is much like that of an alternator and in the case of a machine supplied with three-phase currents, a rotating magnetic flux is produced. The rotor may be either of two basic configurations, which are the "squirrel cage" or the slip-ring type. In the squirrel cage motor the rotor core is laminated and the conductors consist of uninsulated copper, or aluminium, bars driven through the rotor slots. The bars are brazed or welded at each end to rings or plates to produce a completely short-circuited set of conductors. The slip-ring machine has a laminated core and a conventional three-phase winding, similar to the stator, and connected to three slip-rings on the locating shaft. Figure 20.72 shows a schematic representation of an induction motor having three stator coils displaced by 120°.

If the stator coils are supplied with three-phase currents, a rotating magnetic field is produced in the stator. Consider the single rotor coil shown in the figure. At standstill the rotating field will induce a voltage in the rotor coil since there is a rate of change of flux linking the coil. If the coil forms a closed circuit, the induced emf will circulate a current in the coil. The resultant force on the current carrying conductor is a consequence of Eq. (20.17) and this will produce a torque, which will accelerate the rotor. The rotor speed will increase until the electromagnetic torque is balanced by the mechanical load torque. The induction motor will never attain synchronous speed because if it did there would be no relative motion between the rotor coils and the rotating field. Under these circumstances there would be no emf induced in the rotor coils and subsequently no electromagnetic torque. Induction motors, therefore, always run at something less than synchronous speed. The ratio of the difference between the synchronous speed and the rotor speed to the synchronous speed is called the "slip"; i.e.,

$$s = \frac{N_s - N}{N_s} \tag{20.28}$$

The torque–slip characteristic is shown in Fig. 20.73. With the rotor speed equal to the synchronous speed, i.e., $s = 0$, the torque is zero. As the rotor falls below the synchronous speed the torque increases almost linearly to a maximum value dictated by the total of the load torque and that required to overcome the rotor losses. The value of slip at full load varies between 0.02 and 0.06. The induction motor may be regarded as a constant speed machine. The difficulties, in fact, of varying the speed constitute one of the induction motor's main disadvantages.

On start-up, the slip is equal to unity and the starting torque is sufficiently large enough to accelerate the rotor. As the rotor runs up to its full load speed the torque increases in essentially inverse proportion to the slip. The start-up and running curves merge at the full load position.

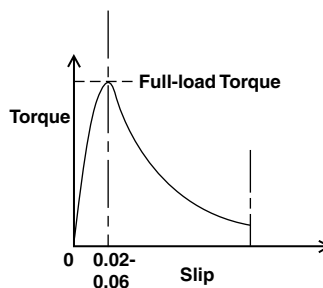



FIGURE 20.73 Torque–slip characteristic for an induction motor.

Starting Induction Motors

As with dc motors, the current drawn during starting of ac motors is very large, up to about five times full load current. A number of devices are therefore employed to limit the starting current but they all involve the use of auxiliary equipment, which is usually quite expensive.

Star-Delta Starter

With the machine at standstill and the starter in the "start" position, the stator coils are connected in the star pattern. As the machine accelerates up to running speed, the switch is quickly moved over to the "run" position, which reconnects the stator windings in the delta pattern. By this simple expedient, the starting supply current is reduced to about one third of what it would have been had the stator windings been connected up in the delta pattern on start-up.

Autotransformer Starter

The autotransformer represents an alternative method of reducing the starting current drawn by an induction motor. The autotransformer incorporates a star connection, which is supplied from a mid-point tapping on each phase. The voltage supplied to the stator is, therefore, one half of the supply voltage. With such an arrangement the supply current and the starting torque are both only one quarter of the values, which would be applied to the motor when the full voltage is supplied. After the motor has accelerated, the starter device is moved to the "run" position thereby connecting the motor directly across the supply and opening the star-connection of the autotransformer. Unfortunately, the starting torque is also reduced and the device is generally expensive since it has to have the same rating as the motor.

Rotor Resistance

With slip-ring induction motors, it is possible to include additional resistance in series with the rotor circuit. The inclusion of extra resistance in the rotor provides for reduced starting current and improved starting torque.

Braking Induction Motors

Induction motors may be brought to a standstill by either "plugging" or by "dynamic braking."

1. **Plugging**: This is a technique where the direction of the rotating magnetic field is reversed. This is brought about by reversing any two of the supply leads to the stator. The current drawn during plugging is very large, and machines which are regularly plugged must be specially rated.
2. **Dynamic braking**: In this braking method the stator is disconnected from the ac supply and reconnected to a dc source. The direct current in the stator produces a stationary unidirectional field and as the rotor will always tend to align itself with the field, it will therefore come to a standstill.

Speed Control of Induction Motors

Under normal circumstances, the running speed of an induction motor will be about 94–98% of the synchronous speed, depending on the load. With the synchronous speed given by Eq. (20.26), it is clear that the speed may be varied either by changing the frequency of the supply current, or by changing the number of poles.

Change of Supply Current Frequency

Solid state variable-frequency drives first began to appear in 1968. They were originally applied to the control of synchronous ac motors in the synthetic fiber industry and rapidly gained acceptance in that particular market. In more recent times they have been used in applications to pumping, synchronized press lines, conveyor lines, and to a lesser extent in the machine-tool industry as spindle drives. Modern ac variable-frequency motors are available in power ratings ranging from 1 to 750 kW and with speed ranges from 10/1 to 100/1.

The synchronous and squirrel cage induction motors are the types most commonly used in conjunction with solid-state, adjustable frequency inverter systems. In operation the motor runs at, or near, the synchronous speed determined by the input current frequency. The torque available at low speed, however, is decreased and the motor may have to be somewhat oversized to ensure adequate performance at the lower speeds. The most advanced systems incorporate a digital tachogenerator to supply a corrective feedback signal which is compared against a reference frequency. This gives a speed regulation of about 3%. Consequently, the ac variable-frequency drive is generally used only for moderate to high power velocity control applications, where a wide range of speed is not required. The comparative simplicity of the ac induction motor is usually sacrificed to the complexity and cost of the control electronics.

Change of Number of Poles

By bringing out the ends of the stator coils to a specially designed switch it becomes possible to change an induction motor from one pole configuration to another. To obtain three different pole numbers, and hence three different speeds, a fairly complex switching device would be required.

Changing the number of poles gives a discrete change in motor speed with little variation in speed over the switched range. For many applications, however, two discrete speeds are all that is required and changing the number of poles is a simple and effective method af achieving this.

Changing the Rotor Resistance

For slip-ring induction motors additional resistance can be coupled in series with the rotor circuit. It has already been stated that this is a common enough method used to limit the starting current of such machines. It can also be used as a method of marginal speed control. Figure 20.74 shows the torque characteristics of a slip-ring induction motor for a range of different resistances connected in series with the rotor windings.

As the external resistance is increased from R_1 to R_3, a corresponding reduction in speed is achieved at any particular torque. The range of speeds is increased at the higher torques.

The method is simple and therefore inexpensive, but the reduction in speed is accompanied with a reduction in overall efficiency. Additionally, with a large resistance in the rotor circuit, i.e., R_3, the speed changes considerably with variations in torque.

Reduced Stator Voltage

By reducing the applied stator voltage a family of torque–speed characteristics are obtained, as shown in Fig. 20.75. It is evident that as the stator voltage is reduced from V_1 to V_3, a change in speed is effected

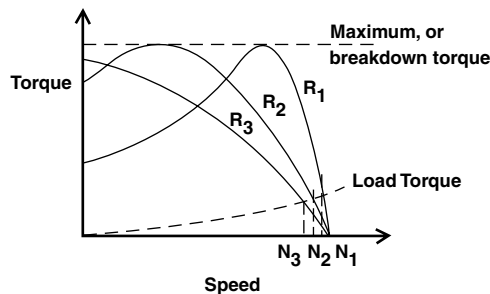



FIGURE 20.74 Torque–speed characteristics for various rotor resistances.

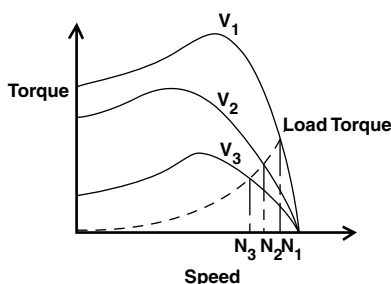



FIGURE 20.75 Torque–speed characteristics for various stator voltages.

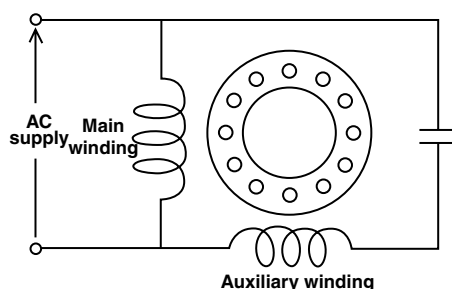



FIGURE 20.76 Capacitor motor.

at any particular value of torque. This is provided, of course, that the torque does not exceed the maximum load torque available at the reduced stator voltage. This latter point is obviously a limiting factor, which places a constraint on this method of speed control. Generally only very small speed ranges can be obtained using variable stator supply voltage.

Single-Phase Induction Motors

The operation of an induction motor depends upon the creation of a rotating magnetic field. A single stator coil cannot achieve this and all of the so-called single-phase induction motors use some or other external means of generating an approximation to a two-phase stator supply. Two stator coils are, therefore, used and these are displaced by 90°. Ideally the currents which supply each coil should have a phase difference of 90°. This then gives the two-phase equivalent of the three-phase induction motor.

The Shaded Pole Motor

The stator of the shaded pole motor consists of a salient pole single-phase winding and the rotor is of the squirrel cage type. One half of the stator features a copper "shading ring." When the exciting coil is supplied with alternating current, the flux produced induces a current in the shading ring. The phase difference between the currents in the exciting coil and the shading ring is relatively small and the rotating field produced is far from ideal. In consequence the shaded pole motor has a poor performance and an equally poor efficiency due to the continuous losses in the shading rings. Shaded pole motors have a low starting torque and are used only in light duty applications such as small fans and blowers or other easily started equipment. Their advantage lies in their simplicity and low cost of manufacture.

The Capacitor Motor

The stator has two windings physically displaced by 90°. A capacitor is connected in series with the auxiliary winding such that the currents in the two windings have a large phase displacement (see Fig. 20.76). The current phase displacement can be made to approach the ideal 90°, and the performance of the capacitor motor closely resembles that of the three-phase induction motor.

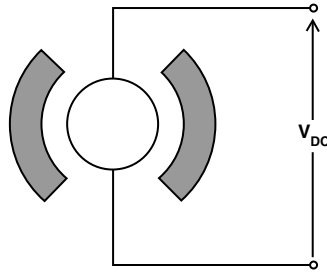



FIGURE 20.77 A dc permanent magnet motor.

The Universal Motor

These are small dc series-wound motors that operate at about the same speed and power on direct current, or on single-phase current with approximately the same root mean square voltage. If alternating current is supplied, the stator and rotor field strengths vary sinusoidally in magnitude but with the same phase relationship. As the applied voltage changes polarity, so do the armature and field currents. Equation (20.25) suggests that under these conditions the applied torque will not reverse polarity and will remain at all times positive. The universal, or plain-series motor, is used mainly in small domestic appliances such as hair dryers, electric drills, vacuum cleaners, hedge trimmers, etc.

The dc Permanent Magnet (PM) Motor

The dc permanent magnet (PM) motor is a continuous rotation electromagnetic actuator that can be directly coupled to its load. Figure 20.77 shows the schematic representation of a PM motor. The PM motor consists of an annular brush ring assembly, a permanent magnet stator ring, and a laminated wound rotor. They are particularly suitable for servo systems where size, weight, power, and response times must be minimized and where high position and rate accuracies are required.

The response times for PM motors are very fast and the torque increases directly with the input current, independently of the speed or the angular position. Multiple pole machines maximize the output torque per watt of rotor power. Commercial PM motors are available in many sizes from 35 mN m at about 25 mm diameter to 13.5 N m at about 3 m diameter.

Direct drive rate and position systems using PM motors utilize dc tachogenerators and position sensors in various forms of closed-loop feedback paths for control purposes.

The Stepper Motor

A stepper motor is a device that converts a dc voltage pulse train into a proportional mechanical rotation of its shaft. In essence, stepper motors are a discrete version of the synchronous motor. The discrete motion of the stepper motor makes it ideally suited for use with a digitally based control system such as a microcontroller. The speed of a stepper motor may be varied by altering the rate of the pulse train input. Thus, if a stepper motor requires 48 pulses to rotate through one complete revolution, then an input signal of 96 pulses per second will cause the motor to rotate at 120 rev/min. The rotation is actually carried out in finite increments of time; however, this is visually indiscernible at all but the lowest speeds.

Stepper motors are capable of driving a 2.2-kW load with stepping rates from 1000 to 20,000 per second in angular increments from 180° down to 0.75°.

There are three basic types of stepper motor, viz.

1. **Variable reluctance**: This type of stepper motor has a soft iron multi-toothed rotor with a wound stator. The number of teeth on the rotor and stator, together with the winding configuration and excitation determines the step angle. This type of stepper motor provides small to medium sized step angles and is capable of operation at high stepping rates.
2. **Permanent magnet**: The rotor used in the PM type stepper motor consists of a circular permanent magnet mounted onto the shaft. PM stepper motors give a large step angle ranging from 45° to 120°.

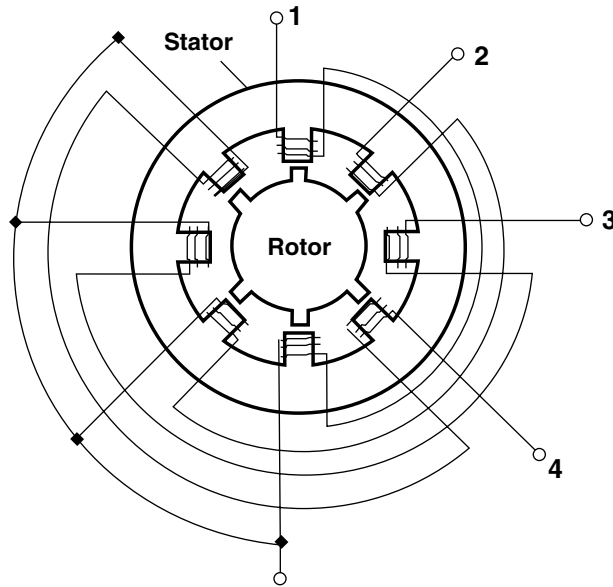



FIGURE 20.78 Variable reluctance stepper motor.

3. **Hybrid**: The hybrid stepper motor is a combination of the previous two types. Typically the stator has eight salient poles, which are energized by a two-phase winding. The rotor consists of a cylindrical magnet, which is axially magnetized. The step angle depends on the method of construction and is generally in the range 0.9°–5°. The most popular step angle is 1.8°.

The principle of operation of a stepper motor can be illustrated with reference to a variable reluctance, four-phase machine. This motor usually has eight stator teeth and six rotor teeth, see Fig. 20.78.

If phase 1 of the stator is activated alone, two diametrically opposite rotor teeth align themselves with the phase 1 teeth of the stator. The next adjacent set of rotor teeth in the clockwise direction are then 15° out of step with those of the stator. Activation of the phase 2 winding on its own, would cause the rotor to rotate a further 15° in the anti-clockwise direction to align the adjacent pair of diametrically opposite rotor teeth. If the stator windings are excited in the sequence 1, 2, 3, 4, then the rotor will move in consecutive 15° steps in the anti-clockwise direction. Reversing the excitation sequence will cause a clockwise rotation of the rotor.

Stepper Motor Terminology

Pull-out torque: The maximum torque that can be applied to a motor, running at a given stepping rate, without losing synchronism.

Pull-in torque: The maximum torque against which a motor will start, at a given pulse rate, and reach synchronism without losing a step.

Dynamic torque: The torque developed by the motor at very slow stepping speeds.

Holding torque: The maximum torque that can be applied to an energized stationary motor without causing spindle rotation.

Pull-out rate: The maximum switching rate at which a motor will remain in synchronism while the switching rate is gradually increased.

Pull-in rate: The maximum switching rate at which a loaded motor can start without losing steps.

Slew range: The range of switching rates between pull-in and pull-out in which a motor will run in synchronism but cannot start or reverse.

The general characteristics of a typical stepper motor are given in Fig. 20.79.

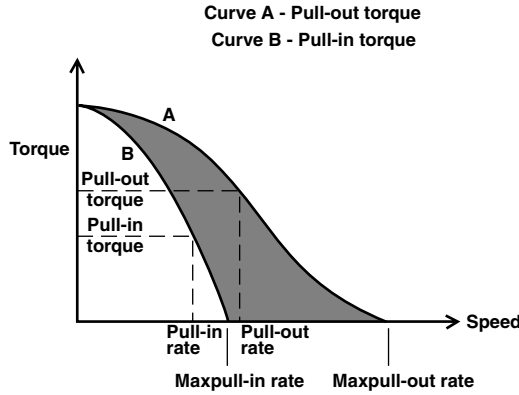



FIGURE 20.79 Stepper motor characteristics.

During the application of each sequential pulse, the rotor of a stepper motor accelerates rapidly towards the new step position. However, on reaching the new position there will be some overshoot and oscillation unless sufficient retarding torque is provided to prevent this happening. These oscillations can cause rotor resonance at certain pulse frequencies resulting in loss of torque, or perhaps even pull-out conditions. As variable reluctance motors have very little inherent damping, they are more susceptible to resonances than either of the permanent magnet, or the hybrid types. Mechanical and electronic dampers are available, which can be used to minimize the adverse effects of rotor resonance. If at all possible, the motor should be selected such that its resonant frequencies are not critical to the application under consideration.

Owing to their unique characteristics, stepper motors are widely used in applications involving positioning, speed control, timing, and synchronized actuation. They are prevalent in *X*-*Y* plotters, floppy disc head drives, printer carriage drives, numerically controlled machine tool slide drives, automatic teller machines, and camera iris control mechanisms.

By far the most severe limitation on the purely electric stepper motor is its power handling capability. Currently this is restricted to about 2.25 kW.

Brushless dc Motors

These motors have position feedback of some kind so that the input waveforms can be kept in the proper timing with respect to the rotor position. Solid-state switching devices are used to control the input signals and the brushless dc motor can be operated at much higher speeds with full torque available at those speeds. The brushless motor can normally be rapidly accelerated from zero to operating speed as a PM motor. On reaching operating speed, the motor can then be switched over to synchronous operation.

The brushless motor system consists of a wound stator, a permanent magnet rotor, a rotor position sensor, and a solid state switching assembly. The wound stator can be made with two or more input phases. Figure 20.80 gives the schematic representation of a two-phase brushless motor.

The torque output of phase A is

$$T_A = I_A(Z\Phi/2\pi)\sin(p\theta/2) = I_A K_T \sin(p\theta/2) \tag{20.29}$$

where

I_A = current in phase A,
$K_T = (Z\Phi/2\pi)$ = torque constant of the motor,
p = number of poles, and
θ = angular position of the rotor.

In the expression for the torque constant, Z is the total number of conductors, and Φ is the magnetic flux.

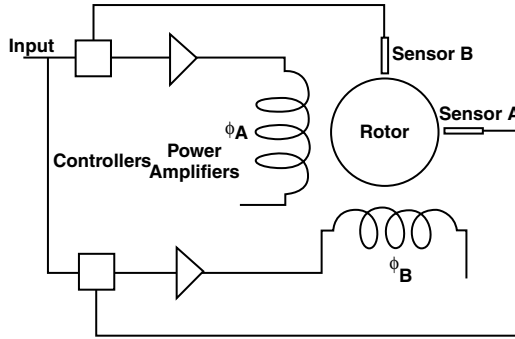



FIGURE 20.80 Two-phase brushless motor.

In a similar manner, the torque output of phase B is

$$T_B = I_B K_T \cos(p\theta/2) \tag{20.30}$$

If the motor currents are arranged to be supplied in the following relationships

$$I_A = I\sin(p\theta/2) \quad \text{and} \quad I_B = I\cos(p\theta/2)$$

then the total torque for a two-pole motor becomes

$$T = T_A + T_B = IK_T[\sin^2(p\theta/2) + \cos^2(p\theta/2)] = IK_T \tag{20.31}$$

Equation (20.31) shows that if all of the above conditions are satisfied then the brushless dc motor operates in a similar manner to the conventional dc motor, i.e., the torque is directly proportional to the armature current. Note that the armature current in this context refers to the stator windings. Excitation of the phases may be implemented with sinusoidal, or square wave inputs. The sine wave drive is the most efficient but the output transistors in the drive electronics must be capable of dissipating more power than that dissipated in square wave operation. Square wave drive offers the added advantage that the drive electronics can be digitally based. The brushless dc motor will duplicate the performance characteristics of a conventional dc motor only if it is properly commutated. Proper commutation involves exciting the stator windings in a sequence that keeps the magnetic field produced by the stator approximately 90 electrical degrees ahead of the rotor field. The brushless dc motor therefore relies heavily on the position feedback system for effective commutation. It might also be apparent that the brushless motor as described is not strictly a dc machine, but a form of ac machine with position feedback.

Motor Selection

For the mechatronics engineer the main concerns regarding electric motors will be those of selection for purpose. At the very least the motor must be capable of matching the power requirements of the driven load. In all cases, therefore, the motor power available should be enough to cope with the anticipated demands of the load. Other requirements are the need for the motor to have enough torque available on start-up to overcome the static friction, accelerate the load up to the working speed, and be able to handle the maximum overload. Too much excess motor torque on start-up might result in a violent initial acceleration. Some systems therefore require a "soft start" whereby the motor torque is gradually increased to allow the load to accelerate gently.

The operating speed of the motor will be fixed by the point at which the torque supplied by the motor is just balanced by the torque requirements of the load. At any other condition, the motor and load will be either accelerating or decelerating. Correct matching of a motor to a driven machine can only be confidently accomplished if both the motor and the load torque–speed characteristics are known. The motor torque–speed characteristics are usually provided by the supplier. The driven machine torque–speed characteristics can be something of an enigma.

Friction devices like industrial sanders, buffers, and polishing machines have a torque–speed characteristic that is initially very high, but drops sharply once motion is established. Continued acceleration usually sees the torque requirement of the load decrease further but at a slower rate than that at start-up. The difference between the static and dynamic friction accounts for this behavior.

Fans and blowers have a torque–speed characteristic that increases parabolically from zero as the speed increases. Such machines do not, therefore, need much motor torque to enable them to start.

High inertia devices like machine tool drives, rolling mills, and electric lifts require a large torque on start-up to overcome the inertia. Once motion is established the torque requirements tend to decrease with increasing speed. The series-wound dc motors are ideal for these types of loads.

This brief discussion of rotating electrical machines is in no way comprehensive. A fuller discourse on ac and dc machines is given both by Gray [1] and Sen [2]. Orthwein [3] presents an interesting practical discussion on the mechanical applications of ac and dc motors and Kenjo & Nagamori [4] provide a detailed in-depth study of permanent-magnet dc motors.

References

1. Gray, C. B. (1989), *Electrical Machines and Drive Systems,* Longmans Scientific and Technical, Harlow.
2. Sen, P. C. (1989), *Principles of Electric Machines and Power Electronics,* Wiley, Chichester.
3. Orthwein, W. (1990), *Machine Component Design,* West Publishing, St Paul, Minnesota.
4. Kenjo, T. & Nagamori, S. (1985), *Permanent Magnet and Brushless dc Motors,* Monographs in Electrical & Electronic Engineering, Clarendon Press, Oxford.

20.3 Piezoelectric Actuators

Ramutis Bansevicius and Rymantas Tadas Tolocka

Piezoeffect Phenomenon

Piezoelectric effect was discovered by brothers Curie in 1880 [1]. The direct piezoelectric effect consists in ability of certain materials to generate electric charge in proportion to externally applied force. The inverse piezoelectric effect of these materials consists in their expansion under the action of electric field parallel to the direction of polarization. Effects are lasting if force or electric field is acting. Effects have been used for actuating/sensing functions in engineering applications.

Constitutive Equations

Coupled electric and mechanical constitutive equations of piezoelectric materials for one dimension medium are

$$S = s^E T + dE \tag{20.31}$$

$$D = \varepsilon^T E + dT \tag{20.32}$$

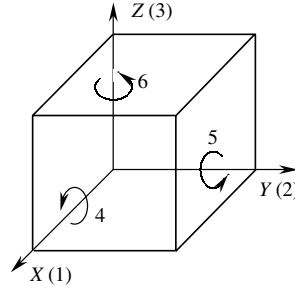



FIGURE 20.81

here S is strain (m), T is stress (N/m^2), E is electric field (V/m), D is dielectric displacement (charge per unit area (C/m^2)), s^E is the compliance of the material (m^2/N) when electric field is constant, ε^T is the permittivity (F/m) under constant stress, and d is piezoelectric constant (m/V or C/N).

First members on the right side of the equations refer to the mechanical properties of an elastic body (Eq. (20.31)) and to electric properties of a dielectric medium (Eq. (20.32)). Artificial piezoelectric materials obtain remnant polarization in the process of poling. Polarization direction coincides with one of the poling electric fields. This direction is referred by convention to the axis Z of orthogonal axes X, Y, Z system. Indexes 1, 2, and 3 are prescribed to these axes, respectively. Piezoelectric properties of piezoelectric materials depend on directions of electrical and mechanical inputs/outputs and are identical along axes 1 and 2. Thus these properties are described by constants with two subscripts, first of which is related to electrical and second to mechanical direction. Subscripts 4, 5, and 6 are used additionally for describing shear distortions in respect to the directions 1, 2, and 3 (Fig. 20.81). Indexes show possible piezo materials operation mode—thickness expansion, transverse expansion, thickness shear, and face shear. The mode of motion depends on the shape and orientation of the body relative to crystal axes and the location of electrodes. Poling electric field direction causes elongation in this direction and contraction in the perpendicular ones. The reverse field causes contraction along the electric field direction and elongation in perpendicular directions. Mode d_{33} gives three times larger displacement than mode d_{31}. Main constants characterizing the piezoeffect are:

- d_{ij} (piezoelectric constant)—strain or charge coefficients expressed in M/V or C/N (according to sensor/actuator piezomaterial properties). They relate to the strain developed by electric field E in the absence of mechanical stress (Eq. (20.31)), and to the electric charge per unit area by the applied stress under zero electric field (Eq. (20.32)). Example: symbol d_{31} means that electrodes are perpendicular to the axis 3 (electric field along it) and stress or strain is along axis 1.
- g_{ij}—voltage or field output coefficients relate to open circuit electric field developed per applied mechanical stress or strain developed per applied charge density and is expressed in V m/N. The relation between d_{ij} and g_{ij} is as follows:

$$g_{ij} = \frac{d_{ij}}{\varepsilon^T} \tag{20.33}$$

- k_{ij} (coupling factors)—energy ratios describing conversion from mechanical to electrical energy and vice versa. Factor k^2 is the ratio of stored converted energy to input energy at operating frequencies far from resonant.

$$k^2 = \frac{d^2}{s^E \varepsilon^T} \tag{20.34}$$

Factor k_p refers to the plane mode operation (strain or stress is equal in all directions perpendicular to axis 3).

TABLE 20.4 Properties of Piezomaterials

Parameter	Units	Pz26	Pz27	Pz29
Relative dielectric constant, K_{33}^{T}	—	1.300	1.800	2.900
Dielectric dissipation factor, tan δ	—	0.003	0.017	0.019
Curie temperature, T_c	°C	330	350	235
Coupling factor, k_p	—	0.57	0.59	0.64
Coupling factor, k_{31}	—	0.33	0.33	0.37
Coupling factor, k_{33}	—	0.68	0.71	0.75
Charge coefficient, d_{31}	m/V	-130×10^{-12}	-170×10^{-12}	-240×10^{-12}
Charge coefficient, d_{33}	m/V	290×10^{-12}	425×10^{-12}	575×10^{-12}
Voltage coefficient, g_{31}	V m/N	-11×10^{-3}	-11×10^{-3}	-10×10^{-3}
Voltage coefficient, g_{33}	V m/N	28×10^{-3}	27×10^{-3}	23×10^{-3}
Frequency constant, N_p	Hz m	2.230	2.010	1.970
Frequency constant, N_t	Hz m	2.040	1.950	1.960
Elastic compliance, s_{11}^{E}	m^2/N	13×10^{-12}	17×10^{-12}	17×10^{-12}
Elastic compliance, s_{33}^{E}	m^2/N	20×10^{-12}	23×10^{-12}	23×10^{-12}
Elastic compliance, s_{11}^{D}	m^2/N	12×10^{-12}	15×10^{-12}	15×10^{-12}
Elastic compliance, s_{33}^{D}	m^2/N	11×10^{-12}	12×10^{-12}	10×10^{-12}
Young modulus	Pa	4.5×10^{-12}	4.5×10^{-12}	4.7×10^{-12}
Poison ratio	—	0.33	0.39	0.34

Other important constants are the Young modulus, relative dielectric constant (ratio of the dielectric permittivity of the material to the dielectric permittivity in vacuum), dielectric dissipation factor (the dielectric loss factor in the material, expressed as tangent of loss angle), Curie temperature (the temperature at which a piezomaterial becomes completely depolarized), etc.

Piezomaterials

Materials that exhibit a significant and useful piezoelectric effect fall into three main groups: natural (quartz, Rochelle salt) and synthetic crystals (lithium sulfate, ammonium dihydrogen phosphate), polarized ferroelectric ceramics, and certain polymer films. The main piezomaterial for engineering applications is ferroelectric ceramics, Lead Zirconate Titanate (PZT) especially. The latter is characterized by high coupling factors, and piezoelectric and dielectric constants over extended temperature and stress ranges. Barium titanate, lead magnesium niobate, modified lead titanate, and bismuth titanate compounds are used as well as other special compositions. Because of their natural asymmetric structure, crystal materials exhibit the effect without further processing. However, ferroelectric ceramics must be artificially polarized by a strong electric field while the material is heated above its Curie point and then slowly cooled with field applied. Remnant polarization being retained, the material exhibits the piezoeffect. Piezopolymers—polyvinylidene fluorides (PVDF or PVF2)—are a special class of fluoropolymer that exhibit a high degree of piezoelectric activity. They are used for manufacturing piezofilms of low thickness (less than 30 μm), which may be laminated on the structural materials.

Table 20.4 shows some properties of typical PZT bulk piezomaterials produced by Ferroperm Piezoceramics [6].

Piezoactuating Elements

Piezoactuating elements in a wide range of sizes are produced as squares, rectangles, rings, discs, spheres, hemispheres, bars and cylinders, and special elements. Typical thickness is from 20 μm to 15 mm and up to 100 mm length or external diameter. Due to unique properties of the material—quick response, high stiffness, low power consumption, and high electromechanical conversion efficiency—they can be used directly as actuators. The piezoeffect is linearly dependant on the applied electric field; its strength in the range 1–2 kV/mm is available and depends on piezomaterial short circuit resistance. Control voltage

TABLE 20.5 Resonant Frequency Expressions for Various Piezoelements

Operating Mode	Expression (L, length; D, diameter; T, width)
Transverse mode, thin bar	$N_1 = f_r \times L$
Radial mode, disc	$N_p = f_r \times D$
Thickness mode, disc	$N_t = f_r \times T$
Length mode, cylinder	$N_3 = f_r \times L$
Shear mode, plate	$N_s = f_r \times T$

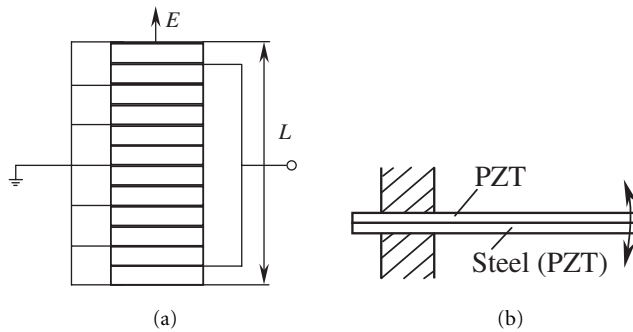
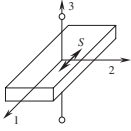
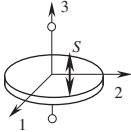
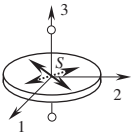
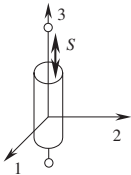
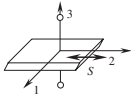



FIGURE 20.82

depends on the thickness of the layer of piezomaterial. Maximum possible relative change in length is up to 0.13%. The shortest time of expansion is one third of the period at resonant frequency oscillations of the mechanical system containing piezoelement. Piezoelement resonant properties are described by frequency constant N_i, which is the resonance frequency f_r multiplied by the linear dimension governing the resonance. Table 20.5 shows the expressions for various operating modes.

Voltage being changed (V remains constant after the change), piezomaterial continues to expand/contract in the same direction, decaying exponentially towards stability. This drift, known as creep, can be estimated by

$$\Delta L(t) = \Delta L(1 + \gamma \lg 0.1) \tag{20.35}$$

where ΔL is 0.1 s expansion after the positioning process, γ is the drift factor. It depends on the design and mechanical load and lies between 0.01 and 0.02. Hysteresis is common in the piezomaterials as well. PZT hysteresis is a fairly constant fraction of the stroke and the width of the hysteresis curve for it can be as large as 20% of the stroke. Due to compensation strategies hysteresis errors decrease up to 3%. Piezoelement with electrodes laminated on to it is electrical capacitor. Because of extremely high piezomaterial internal resistance (more than 100 MΩ) only a small discharge current flows if piezomaterial remains static in the expanded state. Thus, the piezoelement being separated from the source of high voltage, its expansion is decreasing slowly. This, in turn, causes a change in the charge, which results in a current:

$$i = \frac{dQ}{dt} = C\frac{dV}{dt} \tag{20.36}$$

where Q is the charge, C is the capacity, and V is the voltage.

Table 20.6 shows different piezoelement sensing/actuating possibilities for some shape cases. Basic types of piezoactuators used are stacked and of laminar design. Laminar design actuators consist of piezoelectric strips with electrodes bonded onto them. Stacked (multilayer) actuators consist of some thin wafers of piezoactive material between metallic electrodes in parallel connection (Fig. 20.82(a)).

TABLE 20.6 Piezoelement Sensing/Actuators Possibilities

Action Mode (L, length; W, width; T, thickness; D, diameter)	Generated Voltage, V	Displacement, ΔL (ΔT)	Capacitance, C
Transverse length mode: $L > 3W > 3T$	$V = \frac{g_{31}}{W}F$	$\Delta l = \frac{d_{31}L}{T}V$	$C = \frac{\varepsilon_3^T LW}{T}$
Thickness extension mode: $D > 5T$	$V = \frac{4Tg_{33}}{\pi D^2}F$	$\Delta T = d_{33}V$	$C = \frac{\pi\varepsilon_3^T D^2}{4T}$
Radial mode: $D > 5T$	Not applied	$\Delta D = \frac{d_{31}D}{T}V$	$C = \frac{\pi}{4}K_3^T\varepsilon_0\frac{D^2}{T}$
Longitudinal mode: $L > 3D$	$V = \frac{4L}{\pi D^2}g_{33}F$	$\Delta L = d_{33}V$	$C = \frac{\pi D^2}{4L}K_3^T\varepsilon_0$
Thickness shear mode: $W > 5T$, $L > 5T$	$V = \frac{g_{15}}{W}F$	$\Delta x = d_{15}V$	$C = \frac{LW}{T}K_1^T\varepsilon_0$

Note: F is the force and ε_3^T is dielectric permittivity of the material at constant stress in direction 3, K_i^T is relative dielectric constant ($K_i^T = \varepsilon_i^T/\varepsilon_0$), and ε_0 is dielectric permittivity in vacuum.

This way of connection allows greater travel at lower voltage. Usually these wafers are 0.3–1 mm thick. The stack is often referred to operating mode d_{33}. Total travel up to 200 μm can be achieved, and in this case it is in proportion to the number of wafers, if no external load is applied:

$$\Delta l = Vnd_{33} \tag{20.37}$$

where n is the number of elements.

TABLE 20.7 Properties of Some Stacked (Multilayer) Actuators

Standard Part No.	Material	Shape	Length × Width × Thickness (mm) or Ext. Diam. × Intern. Diam. × Thickness (mm)	V_{max} (V)	Stroke (μm)	F_{max} (kN)
A01	Pz26	Rectangle	2.5 × 2.0 × 2.0	200	1.8	0.5
A06	Pz26	Square	10 × 10 × 2.0	200	2.0	10
A16	Pz27	Square	10 × 10 × 2.0	200	3.2	5.0
A21	Pz26	Ring	6.0 × 2.0 × 2.0	200	1.7	2.5
A27	Pz26	Ring	25 × 15 × 2.0	200	2.2	31
A37	Pz27	Ring	25 × 15 × 2.0	200	3.4	16

Due to stacked design, strong pushing force is developed:

$$F = \frac{\Delta l A}{L s_{33} Y} \tag{20.38}$$

where L is the length of the stack, A is the area of elements, and s_{33}, Y are compliance and Young's modulus, respectively.

For reference, Ferroperm Piezoceramics' multilayer actuator supply catalog [6] extract is presented in Table 20.7.

If long travel is required, piezoelement expansion can be amplified by using bimorph or levers. Bimorph is a composite cantilever of two layers (Fig. 20.82(b)). One of them is of structural material and the other of piezomaterial. Piezomaterials can be used for both layers. In this case, first layer will expand, the second one contract. However, this results in low stiffness.

Application Areas

Due to inherent properties in piezomaterials, actuators with a lot of engineering advantages can be developed. Some examples are compact and lightweight, large force, broad operating frequency range, high stability, solid state, displacement proportional to applied voltage, 50% energy conversion efficiency.

They are used in micromanipulation, noise and vibration suppression systems, valves, laser and optics, ultrasonic motors, positioning devices, relays, pumps, in automotive industry, industrial automation systems, telecommunications, computers, etc. Some of the applications are shown in Fig. 20.83.

(a) *Suppression of oscillations.* Piezoactive materials-based dampers convert mechanical oscillations into electrical energy. Generated energy is then shunted to dissipate the energy as heat, i.e., oscillation energy is eliminated. The principle scheme is given in [2].

(b) *Microrobot.* Robot platform legs are piezoactuators. By applying voltage to the electrodes, piezo-legs are lengthened, shortened, or bent in any direction in a fine movement.

(c) *Micropump.* Diaphragm is actuated by piezoactuator, input and output check valves are subsequently opened for liquid or gas pumping. Advantages are fast switching and high compression rate.

(d) *Microgripper.* Piezoactuator works on contraction for gripping motion based on the compliant mechanism. Gripper is of very small size and almost any required geometrical shape.

(e) *Micromanipulator.* Due to the unlimited resolution, piezoactuators are used in numerous positioning applications.

(f) *Microdosage device.* Piezoactuators allow high precision dosage of a wide variety of liquids in a range of nanoliters for various applications.

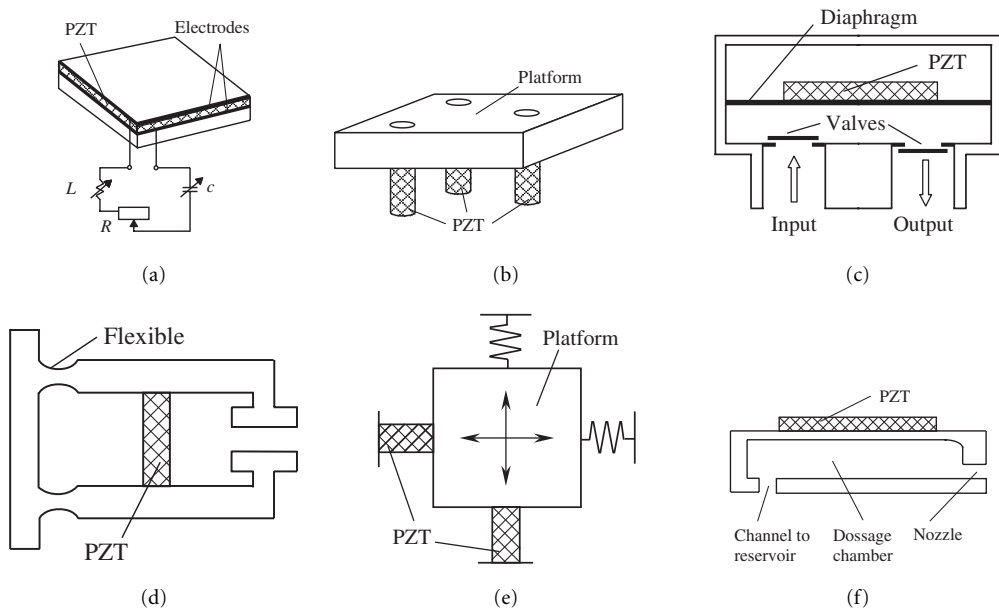
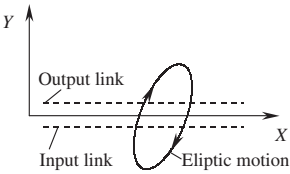
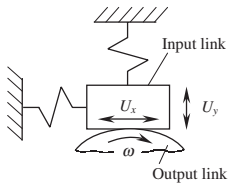
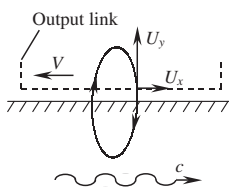
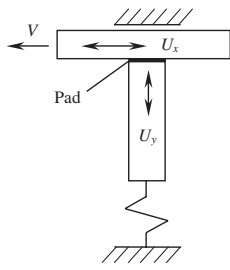
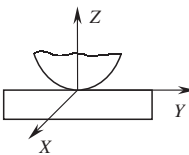
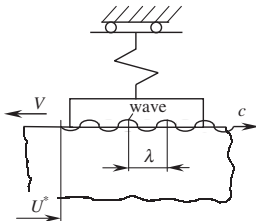
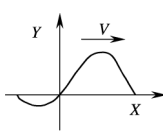
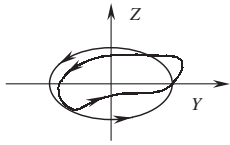
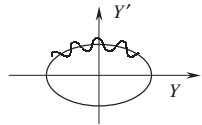
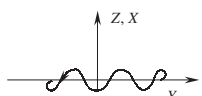
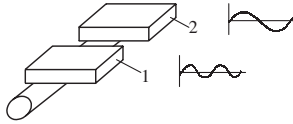
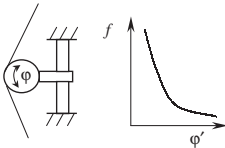



FIGURE 20.83

Piezomotors (Ultrasonic Motors)

Vibromotors [4] find wider and wider application as actuators based on the conversion of high frequency mechanical oscillations (dozens of kHz) into continuous motion [7,8]. Piezoactivating elements can be used as oscillators, and in this case vibromotors are called piezomotors [5,7,8]. Advantages of piezomotors are large torque, high resolution, excellent controllability, small time constant, compactness, high efficiency, silent operation, and no electromagnetic induction.

Main Types of Piezomotors

Piezomotors differ with respect to the methods of oscillations conversion into continuous motion. Basic ideas are given in Table 20.8.

Piezomotors producing elliptical motion in the contact area between input and output links are mainly used. For this purpose oblique impact upon the output link or traveling wave is made use of.

In piezomotors, making use of oblique impact, friction force transmits motion and energy between input and output links. This may be realized by two oscillatory motions (normal and tangential components) u_y and u_x in the contact area with phase difference φ, which is used to change output link motion direction. Both motions can be realized by one or two active links oscillating resonantly. Various oscillations offer possibilities to develop different kinds of piezomotors: longitudinal, transversal, shear, and torsional. Piezomotors employing oblique impacts possess a very wide frequency range. Its lower limit is at lower ultrasound frequencies (for elimination of acoustic action), 16–20 kHz, and its upper limit is at several megahertz. Traveling wave motion piezomotors are based on frictional interaction between the traveling wave motion in the elastic body and the output link, i.e., its principle of operation is similar to the harmonic traction transmission. Wave propagating along the surface (Rayleigh wave) of the input link forms the elliptical motion in the contact area. Rayleigh wave is a coupled wave of longitudinal and shear waves; thus each surface point in elastic medium moves along an elliptical locus. Flexural, shear, torsional, and longitudinal waves are used in piezomotors. Traveling wave in piezoceramic is excited by electrical field.

Traveling wave motion piezomotor characteristics (ABB Corporate Research ITCRC/AS) are shown in the Table 20.9.

TABLE 20.8 Piezomotors Operating Principles

Basic Idea	Schematic of Realization	Remarks
A. Elliptic motion in the contact: two motion components with phase difference 	1. One active link 	$u_y = u_{y0}\sin(\omega t + \varphi)$ $u_x = u_{x0}\sin\omega t$
	2. Two active links 	$u_y = u_{y0}\sin(\omega t + \varphi)$ $u_x = u_{x0}\sin\omega t$ where u_{y0}, u_{x0}, ω, and φ are amplitudes, angular frequency, and phase of oscillatory motions of piezoelements, respectively
B. Elliptic motion in the contact area: traveling wave 		$u = u_0\cos 2\pi/\lambda(u^* - ct)$ where u_0, λ, and c are amplitude, length, and velocity of wave, respectively
C. Frictional anisotropy of contact 	a) b) c) 	Usually $\frac{\tau_c}{T} \geq 0.05$ where τ_c and T are the duration of contact and oscillation period, respectively

TABLE 20.8 Piezomotors Operating Principles (Continued)

Basic Idea	Schematic of Realization	Remarks
D. Asymmetrical oscillations cycles	a) 2 1 b) f φ φ′	

TABLE 20.9 Properties of Some Traveling Wave Piezomotors

Motor	Unit	USR60	USR45	USR30
Operating frequency	kHz	40	43	42
Operating voltage	Vrms	100	100	100
Rated torque	Nm	0.38	0.15	0.04
Rated output	W	4	2.3	1.0
Rated rotational speed	rpm	100	150	250
Mechanical time constant	ms	1	1	1
Weight	g	175	69	33
Rotation irregularity	%	2	2	2
Lifetime	h	1000	1000	1000
Operating temp. range	°C	−10 + 50	−10 + 50	−10 + 50

Traveling wave excitation is achieved simultaneously by excitating different phase oscillations of the same frequency and mode. This is accomplished by dividing the electrodes of the converters into n equal parts and connecting them to the n-phase generator of electrical vibrations, where $n \geq 3$ phases are shifted between adjacent electrodes being $2\pi/n$, or by using discrete converters.

Piezomotors with frictional anisotropy of contact are based on oscillatory motion variations in normal active links contact direction in the cycle of oscillations. This is achieved by superposing additional periodic actions in the contact. The distinguishing feature is time τ_c/T ratio of the reduced duration of the contact to the oscillations period in contact parameter. The contact anisotropy can be achieved in two ways: (a) by locking the active link in a specified segment of the trajectory (Table 20.8, case C, a), (b) by superimposing oscillations of higher frequencies (Table 20.8, case C, b), in the direction of basic oscillations, or in perpendicular direction of basic oscillations (Table 20.8, case C, c) normal or tangential plane.

Piezomotors with asymmetrical oscillations are based on the asymmetry of inertia forces in nonharmonic high frequency oscillations, multiple frequency oscillations (Table 20.8, case D, a), or forces of dry friction with nonlinear relationship between force and velocity (Table 20.8, case D, b). Asymmetric cycles of oscillations are generated by summing harmonics of multiple frequencies. The amplitude of each harmonic is chosen by variating electrode shape and area of divided electrodes or variating amplitude of the voltage supplied. Shift in voltage supply phases is used. Piezomotor efficiency in this case is lower, but designs of devices are characterized by higher—up to 0,002 μm—resolution in translational drive. Besides, this permits piezomotors of limited dimensions in both coordinates, which, in turn, is very important in a number of applications.

Piezomotors are easily miniaturized; thus, micromotors are successfully developed. The rotational motor [5] of this type is a good example. It is 2 mm in diameter, 0.3 mm in height, and its volume is 0.49 mm^3. The motor stably rotates at any posture and the starting torque is about 3.2 μNm.

Piezoactuators with Several Degrees of Freedom

Piezoelectric actuators with several degrees of freedom allow new class of mechanisms, capable of changing their parameters or kinematic structure under control. If one or both links of the kinematic pair are made from piezoactive material, it is possible to generate static displacement of its elements and quasi-static or resonant oscillations, resulting in generating forces or torque in contact area of links. Motion of one link relative to the other is obtained. Such kinematic pairs can be defined as active. Active kinematic pairs are characterized by

- Control of number of degrees of freedom. The simplest one is to control friction in the pair, usually when the elements of the pair are closed by force. Here either the friction coefficient or magnitude of the force executing the closure can be varied. This is achieved by excitation of high frequency tangential or normal vibrations in the contact area of the pair.
- Generation of forces or torque in the contact area between links. The direction of generated forces or torque is controlled by special shift of oscillations, e.g., by activating specific by sectioned electrodes of the transducer.
- Possibilities to realize additional features: self-diagnostics, multifunctionality, self-repair, self-adaptation.

The example is a robot's eye (Fig. 20.84(a)) in which miniature CCD camera 1 is fixed in the passive sphere 2, contacting with piezoelectric ring 3. Constant pressure in the contact zone is realized by

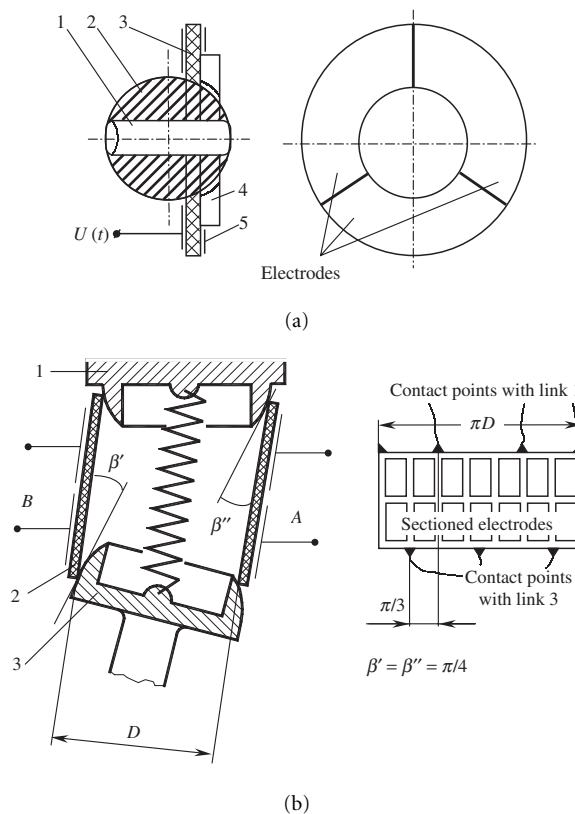



FIGURE 20.84

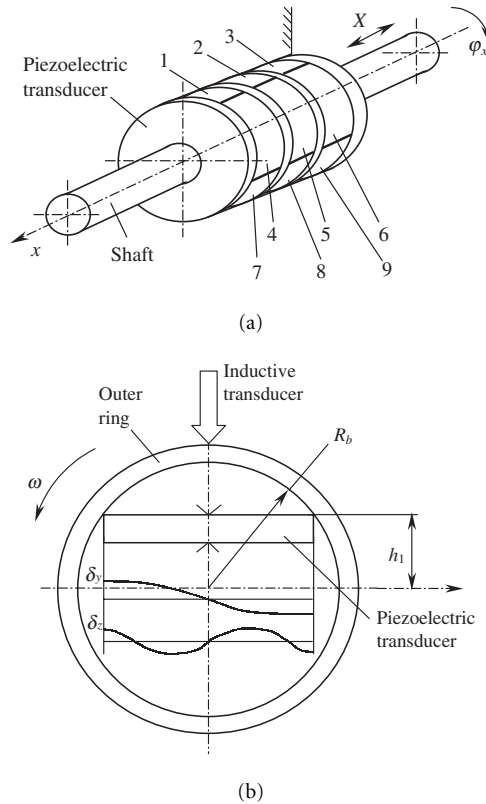



FIGURE 20.85

permanent magnet 4. The system is a kinematic pair possessing three degrees of freedom. Electrodes 5 in the piezoelectric ring are sectioned (in this case - into three symmetric parts); activating any of them with AC of resonant frequency results in the rotation of the sphere around its axis, position of which is controlled by changing the activated electrode. Traveling wave oscillations generated in the ring (by applying three-phase AC to all three electrodes) result in the rotation of the sphere around the axis of the ring. Such type actuators possess approximately two angular seconds resolution in every direction—higher than the requirements for robot vision systems.

The existing methods to control types and forms of resonant oscillations make it possible to design mechanisms with the same active link being used in two kinematic pairs to increase redundancy in the system. A piezoelectric robot, based on active kinematic pairs, is shown in the Figure 20.84(b). It consists of passive material (e.g., steel) spheres 1 and 3, with piezoelectric transducer 2 between them. Springs ensure contact between all links. A robot with two spherical kinematic pairs possess six degrees of freedom. Kinematic pairs move due to piezomotor design methods. The electrodes on active links are sectioned. Figure 20.84(b) shows their form and the distribution of the three component oscillations in the contact area. High frequency multicomponent oscillations generated at the contact points (certain electrodes “A” of link 2 are actuated) rotate link 2 in relation to link 1. A $\pi/3$ change in the position of oscillation pattern (a change of position of vibration nodes in contact points) results in the rotation of link 3 in relation to link 2. Using direct piezoeffect, it is possible to extract additional information (with the help of electrodes “B”) on forces and torque, acting on link 2 and on the state of contacting surfaces. This information is used to reduce positioning errors and to correct motion trajectory.

Classically, by increasing accuracy and stiffness of system elements, static and dynamic errors in bearings, supports, and guides are decreased or eliminated completely. By integrating unique properties of piezoactive transducers and actuators in the control system it is possible to sharply reduce or even fully

eliminate most errors in bearings, supports, and guides used in high precision measuring devices. This is due to active bearings and supports possessing several degrees of freedom, in which one or both contacting elements are made from piezoactive material with predetermined excitation zones. Radial or axial play, backlash and dead zones—traditional errors—are minimized in these devices. The schematic of active bearing is shown in Fig. 20.85(a) where number of axial n and radial m electrode sectors is $n = m = 3$. Active bearings are used in precision component surface and profile measuring systems to scan the surface. The example is outer ring errors evaluation in high precision ball bearings. Here rotating the component simultaneous measurements of profile and surface are obtained. This is possible due to piezoelectric transducers (Fig. 20.85(b)) contacting with the component in two areas with the same pattern of oscillation distribution and phase shift between normal and tangential components of oscillations. There being no external forces, it is evident that errors caused by torque, generated in the contact zone, are negligible.

References

1. Cady, W. G., *Piezoelectricity,* Dover Publications, New York, 1964.
2. Volkov, V., Some Theoretical Problems in Modern Techniques of Diagnostics in Mechanical Systems, in *Proc. Int. AMSE Conf. Systems Analysis, Control and Design,* Lyon, France, 205.
3. Uchino, K., *Piezoelectric Actuators and Ultrasonic Motors,* Kluwer Academic Publishers, MA, 1997, 349.
4. Ragulskis, K., Bansevicius, R., Barauskas, R., Kulvietis, G., *Vibromotors for Precision Microrobots,* Hemisphere Publishing Corporation, 1988, 310.
5. Suzuki, Y., Tani, K., Sakuhara, T., Development of new type piezoelectric micromotor, *J. Sensors & Actuators,* 83, 244, 2000.
6. Catalog Ceramic Multilayer Actuator CMA d_{33} & d_{31}, July 2000.
7. Sashida, T., Kenjo T., *An Introduction to Ultrasonic Motors,* Oxford Science Publications, 1993, Oxford University Press, New York, 242.
8. Ueha S., Tomikawa Y., *Ultrasonic Motors, Theory and Application,* Oxford Science Publications, Oxford Press, Oxford, 1993, 298.

20.4 Hydraulic and Pneumatic Actuation Systems

Massimo Sorli and Stefano Pastorelli

Introduction

The primary function of an actuation system is to influence the controlled system so as to obtain the desired movement or action. This objective is made possible by the actuation system, which converts the primary energy with which the actuator operates into the final mechanical energy.

There are three main types of power with which actuation systems work: electric power, hydraulic power, and pneumatic power. The first envisages the use of electric actuators such as motors, solenoids, and electromagnets. The remaining two envisage the use of cylinders (linear motors) and rotary motors, substantially similar in form and dimensions, the motion of which is respectively governed by a fluid considered uncompressible in an initial approximation (a hydraulic liquid, mineral oil generally, or a liquid with lower viscosity) and by a compressible fluid (compressed air or a generic gas).

Other types of energy are available but are fairly unusual in automatic systems. Chemical energy and thermal energy, which cause a change of phase in a material or the thermodynamic expansion of the systems into a mechanical movement, can be considered in this category.

The characteristics of fluid servosystems are examined below, with particular reference to systems which permit continuous control of one of the two physical magnitudes which express the fluid power: pressure and flow rate. In general, pressure control is carried out in cases in which it is necessary to create a determined force or torque law, while flow rate control is used to carry out controls on kinematic magnitudes such as position, speed, and acceleration.

Continuous control of a force or of a speed can be effectively realized with a fluid actuation device, with evident advantages compared with electric actuation, such as the possibility of maintaining the system under load without any limitation and with the aid of adequate control devices, the possibility of carrying out linear movements directly at high speeds, without devices for transforming rotary motion to linear, and the possibility of having high bandwidths, in particular in hydraulic systems, as these have limited dimensions and therefore low inertia.

Fluid Actuation Systems

An actuation system, which is part of an automatic machine, consists of a power part and a control part as illustrated in Fig. 20.86. The power part comprises all the devices for effecting the movements or actions. The control part provides for the processing of the information and generates the automated cycle and the laws of variation of the reference signals, in accordance with the governing procedures implemented and with the enabling and feedback signals arriving from the sensors deployed on the operative part. The order signals coming from the control part are sent to the operative part by means of the interface devices which convert and amplify the signals, where necessary, so that they can be used directly by the actuators. These interfaces can be the speed drives or the contactors of the electric motors, the distributor valves in hydraulic and pneumatic actuators.

Figure 20.87 illustrates a fluid actuation system. The power part consists of the actuator—a double-acting cylinder in the case in the figure—the front and rear chambers of which are fed by a 4/2 distributor valve, which constitutes the fluid power adjustment interface.

The valve switching command is the order from the control part. This order is sent in accordance with the movement strategy, determined by the desired operating cycle of the cylinder in the control part, on the basis of the feedback signals from the sensors in the cylinder, represented in the figure by the limit switches.

Then there are discontinuous actuation systems and continuous actuation systems, depending on the type of automation realized, while retaining the control part and the actuation part. The first are effective when used in discontinuous automation, typical of assembly lines and lines for the alternating handling of machine parts or components; on the other hand, continuous actuation systems are found in continuous process plants and as continuous or analog control devices for the desired magnitudes, and constitute fluid servosystems.

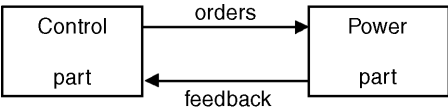



FIGURE 20.86 Actuation system.

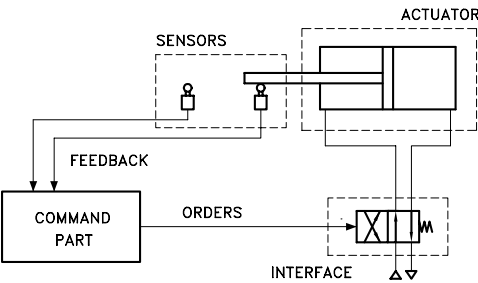



FIGURE 20.87 Fluid power actuation system.

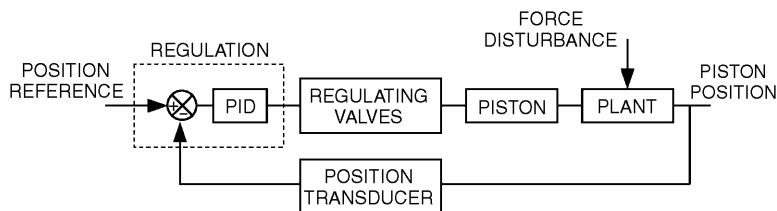



FIGURE 20.88 Scheme of a fluid power servosystem.

Fluid actuators, whether they are linear (cylinders) or rotary (motors) are continuous systems as they can determine the positioning of the mobile component (of the rod with respect to the cylinder liner; of the shaft with respect to the motor casing) at any point in the stroke. Performance of the usual cylinders and motors is currently highly influenced by the action of friction (static and dynamic) developed by contacts between mobile parts. This action, in pneumatic systems in particular, gives rise to the well-known phenomenon of stick-slip, or intermittent motion at very low movement speeds, due to the alternation of conditions of friction and adherence in the motion of the mobile element in the actuator. Given the nature of the friction itself, the presence of devices suitable for sustaining the mobile components of the actuator and maintaining the correct pressure conditions, such as supports and gaskets, gives rise to nonlinear conditions in the equilibrium of the actuator, increasing the level of difficulty in obtaining high precision in positioning the system. To overcome these problems in specific applications it is necessary to use actuators without seals, for example, with fluid static and/or fluid dynamic bearings.

The interface element, indicated as a distributor in the figure, takes on a crucial role in the definition of the operating mode of the actuator. Indeed, in the case in which it is only necessary to create reciprocating movements, with positioning of the actuator at the end of its stroke, it is only necessary to use a two- or three-position distributor valve, with digital operation. This is the solution shown in Fig. 20.87.

If, on the other hand, it is necessary to have continuous control of the position and force transmitted, it is necessary to use devices which are not digital now, but which are continuous, such as proportional valves and servovalves, or it is necessary to use digital devices operating with control signal modulation, for example those of the PWM (Pulse Width Modulation) type.

The actuation system therefore becomes a fluid servosystem, such as the one outlined in Fig. 20.88, for example. A practical construction of a hydraulic linear servoactuator having the same working scheme of Fig. 20.88 is shown in Fig. 20.89. It consists of a cylinder, a valve, and a position transducer integrated in a single device.

A controlled, fluid-actuated system is a classical mechatronic system, as it combines mechanical and fluid components, and control and sensing devices, and normally requires a simulation period for defining the size and characteristics of the various elements so as to comply with the desired specifications.

The standardized symbols for the different components of hydraulic and pneumatic fluid systems, and the definitions of the associated circuits, are defined in the standard, ISO 1219 “Fluid power systems and components—Graphic symbols and circuit diagrams; Part 1: Graphics symbols, Part 2: Circuit diagrams.”

Fluid Servosystems

Fluid servosystems are devices for controlling a generically mechanical output power, either by controlling a kinematic magnitude (servosystems for controlling position or speed) or by controlling an action (servosystems for controlling the force, torque, or pressure).

The output magnitude control action is obtained by controlling the fluid power, that is, by the power of the fluid passing through the components of the servosystem.

Two large classes of fluid servosystems are usually present in current applications: hydraulic servosystems, in which the operating fluid is a liquid, and pneumatic servosystems, in which the fluid used is compressed air. The working pressure in hydraulic servosystems is typically comprised between 150 and 300 bar, while in the case of pneumatic systems, the pressure values are generally below 10 bar.

FIGURE 20.89 Hydraulic servocylinder (Hanchen).

The first group obviously includes hydraulic oils, that is, fluid with high viscosity, now traditionally used in servosystems in which a high controlled pressure is requested, but also combustible fluids, such as automotive or aeronautical petrols (JPA, JPB,…), used in all the applications found in the fuel circuits of combustion engines. Other servosystems include those which use both industrial and seawater as the working fluid. The latter solution has unquestionable advantages in all naval and off-shore applications.

Pneumatic servosystems include all the industrial applications for automation of production and process automation, and also the vehicular applications on means of air, sea, road, and rail transport. The compressed air in these applications is generated by compressors using air drawn in from the environment. Further applications include those in which the working fluid is not compressed air but a particular gas. In this regard, there are servosystems with refrigerant fluids in the gaseous stage, in both vehicular and industrial cryogenic systems, with fuel gases (LPG, methane, propane) in domestic applications, and with nitrogen in high-pressure applications.

It can be seen from this preliminary analysis that fluid servosystems are present both in the realization of a product, being integral parts of the automated production process, along with the electric servomechanisms, and as controlled actuation devices integrated in the product itself; in this regard we can mention generic servoactuators installed on aeroplanes and increasingly in road vehicles today.

Hydraulic Actuation Systems

The components of a hydraulic actuation system are:

- the pump, that is, the hydraulic power generation system;
- the actuator, that is, the element which converts hydraulic power into mechanical power;
- the valve, that is, the hydraulic power regulator;
- the pipes for connecting the various components of the actuation system;
- the filters, accumulators, and reservoirs;
- the fluid, which transfers the power between the various circuit elements;
- the sensors and transducers;
- the system display, measurement, and control devices.

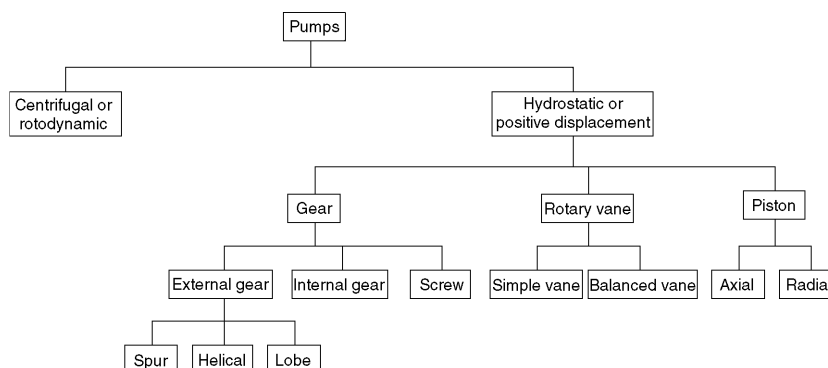



FIGURE 20.90 Pumps classification.

Pumps

Pumps transform electrical or mechanical energy into hydraulic energy. They constitute the fluid flow generator of the hydraulic system, as the pressure is determined by the fluid resistance downstream from the generator. The main types of pumps are shown in Fig. 20.90.

Centrifugal pumps permit high deliveries with low pressures. They do not have internal valves but have a large clearance between the rotary part and stator part and guarantee a sufficiently stationary flow. Vice versa, hydrostatic or positive displacement pumps, which are those most commonly used, guarantee high pressures with limited deliveries. They have elements such as valves and caps, which permit separation of the delivery zone from the intake zone, and they may introduce pulses in the flow in the delivery line and generally require the use of a fluid with sufficient lubricating properties and load capacity, so as to reduce the friction between the sliding parts of the pump. There are constant displacement and variable displacement pumps.

The main positive displacement pumps belong to the gear, rotary vane, and piston types.

Gear Pumps

Gear pumps are subdivided into pumps with external gears, pumps with internal gears, and screw pumps. In all cases, the pump is made up of two toothed wheels inserted into a casing with little slack so as to minimize leakage.

Figure 20.91 is a photograph of a pump with external gears. The opposed rotation of the wheels causes the transfer of the oil trapped in the space between the teeth and walls of the gear from the intake to the outlet. Depending on the form of the teeth, there are external gear pumps of the spur gear, helical gear, and lobe gear types.

Pumps with internal gears are functionally similar to the above, but in this case the gears rotate in the same direction. Figure 20.92 is a section plane of a two-stage pump. In screw pumps, which may have one or more rotors, the elements have helical toothing similar to a threaded worm screw. Transfer of the fluid takes place in an axial direction following rotation of the screw. These types of pump guarantee very smooth transfer of the flow, with reduced pulsation and low noise levels.

The usual rotation speeds are between 1000 and 3000 rpm, with powers between 1 and 100 kW. Delivery pressures can reach 250 bar, with higher values in the case of the pumps with external gears. The flow transferred is a function of the pump displacement and the angular input speed, with values comprised between 0.1 and 1000 cm^2/rev. Double pumps can be used to increase these values. Gear pumps have high performance levels, with values around 90%.

Rotary Vane Pumps

Vane pumps (Fig. 20.93) generally consist of a stator and a rotor, which can rotate eccentrically with respect to one another. Vanes can move in special slits placed radially in the stator or in the rotor and

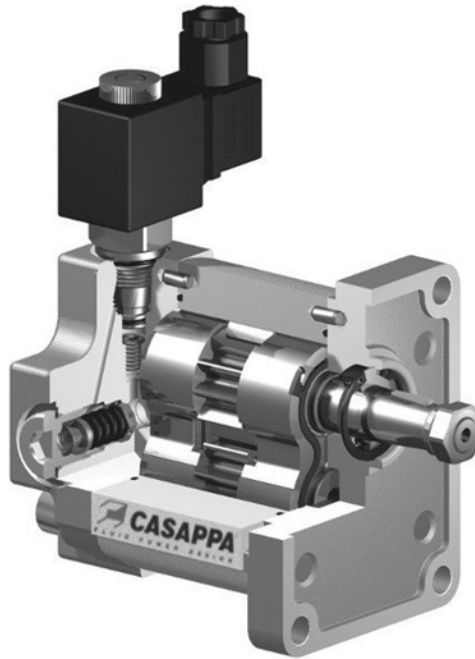



FIGURE 20.91 External spur gear pump (Casappa).

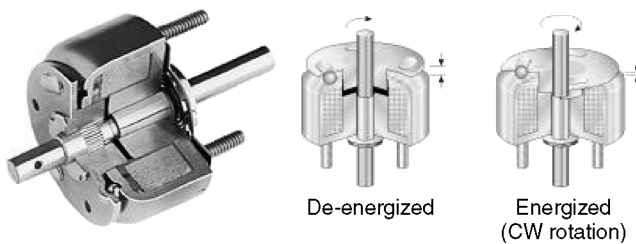



FIGURE 20.92 Internal gear pump (Truninger).

delimit appropriate variable volumes. In Fig. 20.93, as in most constructions, the vanes are borne by the rotor which can rotate inside the stator. Rotation leads to the displacement of volumes of fluid enclosed between two consecutive vanes from the intake environment to input into the delivery environment. This type of pump permits a range of working pressures up to 100 bar and, compared with gear pumps, guarantees lower pulsing of the delivery flow and greater silence.

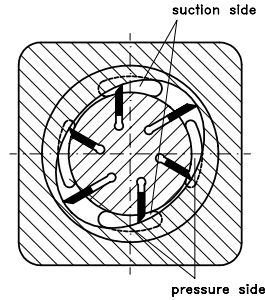



FIGURE 20.93 Rotary vane pump.

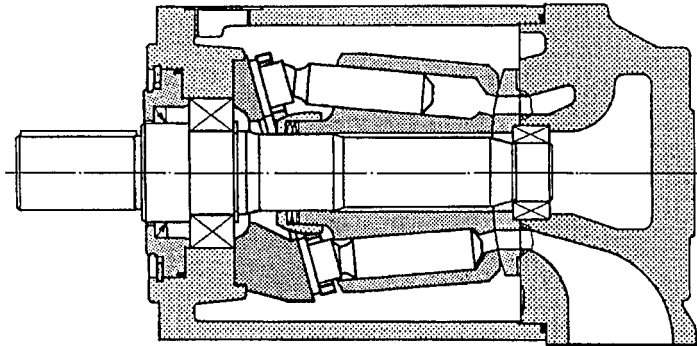



FIGURE 20.94 Axial piston swash plate pump (Bosch Rexroth).

Piston Pumps

Volumetric piston pumps can have one or more cylinders; that is, there may be one or more cylinders with a piston sliding in each of them. Transfer of the volume of fluid from intake to delivery is determined by the displacement of the piston inside the cylinder, which is provided with input and output valves or shutters. Depending on the geometrical arrangement of the cylinders with respect to the rotating motor shaft, piston pumps are subdivided into axial pumps (bent axis type and swash plate type) and radial pumps. Figure 20.94 shows the plan of a fixed-displacement axial piston pump, of the swash plate type. The working pressure range available with the aid of piston pumps is greater than in the previous cases, being able to reach pressures in the order of 400–500 bar but with the disadvantage of more uneven flow.

Motion Actuators

Motion actuators convert the hydraulic energy of the liquid under pressure into mechanical energy. These actuators are therefore volumetric hydraulic motors and are distinguished, on the basis of the type of movement generated, similar to what has been said about pumps, into rotary motors, semi-rotary motors or oscillating ones, which produce limited rotation by the output shaft, and into linear reciprocating motors, that is hydraulic cylinders.

Rotary and Semi-rotary Motors

In construction terms, rotary motors are identical to rotary pumps. Therefore gear, vane, and piston motors, radial or axial, are available. Obviously, the operating principle is the opposite of what has been said for pumps. The symbols of hydraulic rotary motors are shown in Fig. 20.95. Semi-rotary motors generate the oscillating motion either directly, by means of the rotation of a vane connected to the output shaft, or indirectly, by coupling with a rack, driven by a piston, with a toothed wheel connected to the output shaft, as in the example in Fig. 20.96. The semi-rotary vane motors produce high instantaneous torsional torque on the output shaft; for this reason they are also called hydraulic torque-motors.

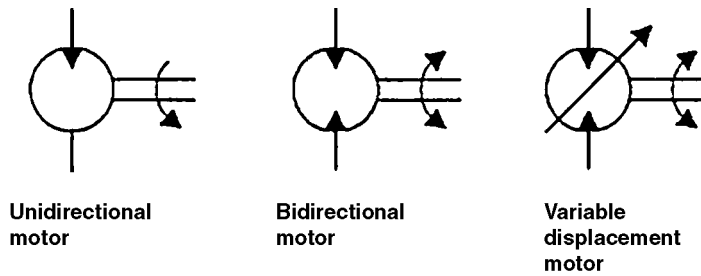



FIGURE 20.95 Symbols of hydraulic rotary motors.

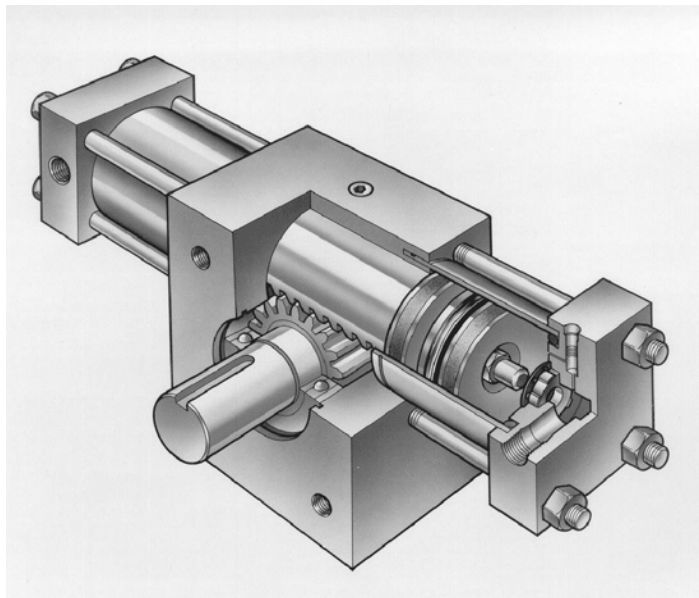



FIGURE 20.96 Hydraulic rotary actuator (Parker Hannifin).

Linear Actuators

Linear hydraulic motors constitute the most commonly used type of actuator. They provide a rectilinear movement realized by the stroke of a rod connected to a piston sliding inside the cylinder. A distinction is made between single acting and double acting cylinders. The former only permit a single work stroke and therefore the pressure of the fluid is exerted on the surface of the piston in one single direction; the retract stroke is made by means of the force applied externally to the cylinder rod, or with the aid of a helical spring incorporated with the actuator inside a chamber. The latter permit both strokes, so that the fluid acts alternately on both faces of the piston, generating both the advance and retract strokes. Double acting cylinders may have a single rod or a double through rod. These are composed of a tube closed at the ends by two heads, and a mobile piston inside the barrel bearing one or two rods connected externally to the load to move. As it is fitted with sealing gaskets, the piston divides the cylinder into two chambers. By sending the oil under pressure into one of the chambers through special pipes in the heads, a pressure difference is generated between the two surfaces of the piston and a thrust transmitted to the outside by the rod. Figure 20.97 shows the constructional solution of a hydraulic double acting cylinder with a single rod. Single rod actuators are also known as asymmetrical cylinders because the working area on the rod side is smaller than the area of the piston, as it is reduced by the section of the rod itself.

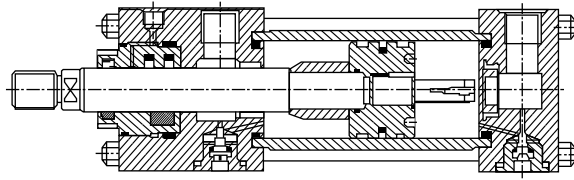



FIGURE 20.97 Single rod double-acting piston actuator (Atos).

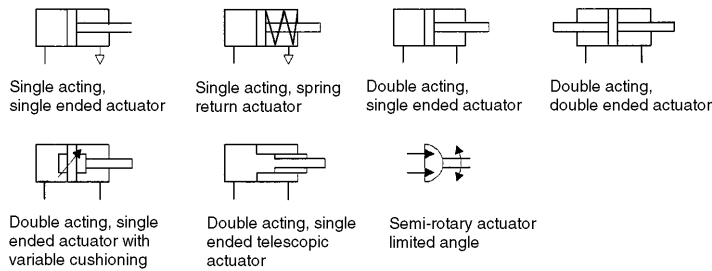



FIGURE 20.98 Actuators simbols.

This involves actuating forces and feed speeds which are different in the two directions, with the same feed pressure in the two thrust chambers.

Hydraulic actuators are able to support external overloads, as, if the load exceeds the available thrust force, the rod stops or reverses, but generally does not suffer any damage. Cylinders may get damaged however, or at least suffer a drop in performance, when they have to support loads which are not applied along the axis of the rod, that is, with components in the radial direction, as reactions are generated on the rod supports and piston bearings, which leads to fast wear of the same and reduces the tightness with oil leakage as a result.

The main features of a linear actuator are its bore, its stroke, its maximum working pressure, the type of working fluid, and the way its connections are fitted.

The symbols of the different types of actuators can be seen in Fig. 20.98.

Valves

Valves are the components in hydraulic circuits that carry out the task of regulating the hydraulic power sent to the actuator. Their role is to turn the oil flow on or off or to divert it according to needs, thereby permitting adjustment of the two fundamental physical magnitudes of fluid transmission: pressure and flow rate. They are subdivided as follows on the basis of the operations they carry out:

- directional valves
- on-off valves
- pressure regulator valves
- flow-rate regulator valves

In servomechanism applications valves with the continuous positioning of the moving components in them, said flow proportional valves or servovalves, and pressure proportional valves are used.

Directional Valves

Directional valves determine the passage and the flow direction of the oil current by means of the movement of appropriate moving parts contained in them, actuated from outside. Directional valves,

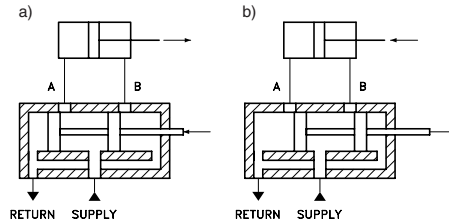



FIGURE 20.99 Scheme of four-way two-position valve.

also known as distributors, are distinguished according to the type of mobile element and therefore of their internal structure, by the number of possible connections with external pipes and by the number of switching positions.

The mobile element can be a poppet type or a spool type. Poppet valves are indifferent to fluid type and are not affected by impurities in the fluid, but require high actuating forces as it is not possible to compensate for the hydraulic forces of the oil pressure. Spool valves permit simultaneous connection to several ways and different switching schemes and therefore are more common because of their variability. The number of possible connections is defined by the number of hydraulic connections or ways present on the external body of the valve. The number of switching positions corresponds to the number of connection schemes which a valve makes it possible to obtain by means of appropriate movements of the mobile element.

Figure 20.99 shows the operating scheme of a four-way, two-position spool valve (indicated as 4/2) connected to a double acting linear actuator. In the first position (Fig. 20.99(a)) the supply is in communication through output A with the rear chamber of the cylinder, while the front chamber discharges through port B. In this configuration, the piston effects an advance stroke with the rod coming out. In the second position, (Fig. 20.99(b)), the result of the movement of the slide valve is that the feed and discharge conditions of the two chambers are inverted, and therefore, a retract stroke is effected.

A directional valve with several positions is represented symbolically by means of quadrants side by side depicting the connections made by each position. Figure 20.100, for example, shows some directional valve symbols in accordance with ISO standards. The central configuration of the three-position valves, which is normally the rest position, is linked with the geometry of the valve spool and of the associated seats.

Directional valves can be controlled in various ways (Fig. 20.100): manually, by applying muscle power; mechanically, by means of devices such as cams, levers, etc.; hydraulically and pneumatically, by means of fluids under pressure; and electromagnetically, directly or piloted, depending on whether the positioning force is generated directly by the electromagnet placed in line with the slide valve, or by means of a hydraulic fluid, the direction of which is managed by a pilot valve which is smaller than the main controlled valve.

On-Off Valves

On-off valves are unidirectional valves, which permit the fluid to flow in one direction only. Because they impede flow in the opposite direction they are also called nonreturn or check valves. On-off valves are normally placed in the hydraulic circuit between the pump and the actuator so that, when the generator stops, the fluid contained in the system is not discharged into the reservoir but remains in the piping. This prevents a waste of energy for subsequent refilling and guarantees positioning of the actuator under load.

Constructively, check valves consist of an actuator, with ball or piston, which in the impeded flow configuration is maintained in contact against its seat by the thrust of a spring (nonreturn valve), or by the pressure difference between inlet and outlet (unidirectional valve).

Pressure Regulator Valves

There are essentially two types of pressure regulator valves: pressure limiter valves or relief valves, and pressure reduction valves.

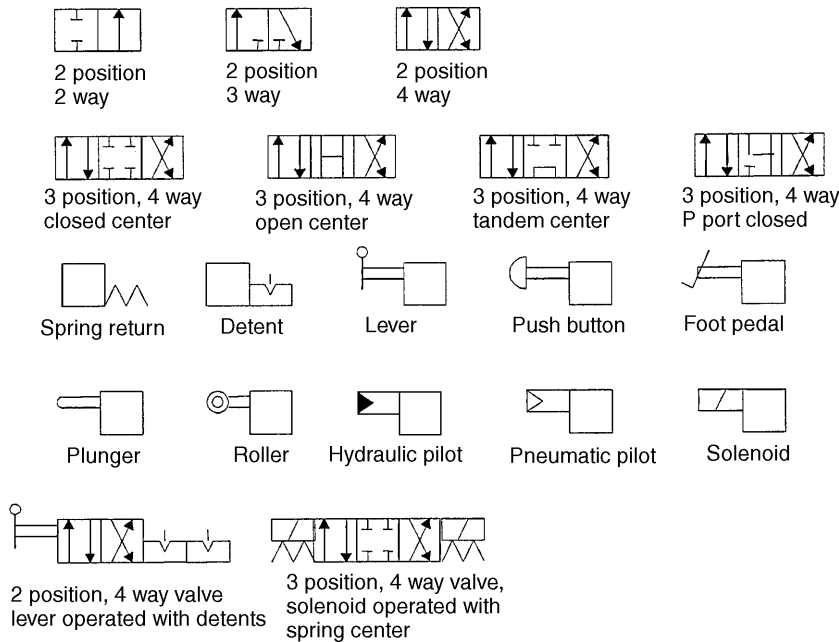



FIGURE 20.100 Valves symbols.

Relief valves guarantee correct operation of the system, preventing the pressure from exceeding danger levels in the system itself. There is always one maximum pressure valve in a hydraulic circuit to discharge any excess flow not used by the system back towards the reservoir. This is because the generator, or positive-displacement pump, provides a continuous flow of fluid which, if not absorbed by the user and in the absence of a relief or maximum pressure valve, would let the pressure in the system increase to unacceptable values. Pressure limiter valves can be direct-acting or piloted. The first provides the force of a spring with a fixed preload as the force contrasting the pressure of an obturator or an adjustable one, which guarantees the maximum opening pressure. The latter replaces the action of the spring with that of the hydraulic control fluid managed by a pilot valve.

The function of the pressure regulator valves is to maintain a constant pressure valve downstream from them, independently from variations in the upstream pressure. The regulated pressure value can be set manually, by means of a pilot signal, or by an electrical analog command. In the latter case, pressure regulator valves may operate in closed electrical loops, as they have an internal transducer to measure the controlled pressure.

Flow-rate Regulator Valves

A flow-rate regulator valve makes it possible to control the intensity of the flow of fluid passing through it. Functionally it operates as a simple restriction, similar to an orifice, with a variable area. The flow passing through a restriction is a function of the area of passage and of the difference in the pressures upstream and downstream from the component. The simple restriction is therefore sensitive to the load, as the flow rate also depends on the pressure drop at its ends, which is established by the other components in the circuit.

In the case of a pressure-compensated flow regulator valve, the flow rate is found to be maintained sufficiently constant above a minimum pressure stage (typically 10 bar) as an exclusive function of the external manual or electrical set-point. In this case, the valve has two restrictions in series, one of which is fixed and the other automatically variable, so as to maintain the pressure drop constant on the fixed restriction and guarantee the constancy of the flow rate.

The symbols for flow regulator valves in accordance with ISO standards are given in Fig. 20.101.

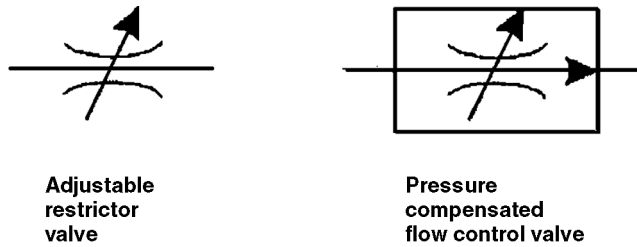



FIGURE 20.101 Symbols of flow control valves.

Proportional Valves and Servovalves

Servovalves began to appear at the end of the 1930s and were mainly used in the military and aeronautical fields. The first commercial versions appeared in the mid-50s. Servovalves and proportional valves are widely used today in the civil field, in the aeronautical, aerospace, automotive, and industrial sectors. In general, they are used for the continuous control of the displacement, speed, and force of a hydraulic actuator from which high performance is requested in terms of positioning precision, or accuracy in up and running conditions, and of working frequency bandwidth amplitude, both in open and closed loop control configurations.

A servovalve or proportional valve is a fluid component capable of producing a controlled output as a function of an input of electrical type. The device converting the electric signal into an action of the spool or poppet of the valve is electromagnetic, of the torque motor or proportional solenoid type. The torque motor converts a small DC current into torque acting on the rotor plate, in bipolar mode. Proportional solenoids produce a unidirectional force on the mobile armature function of the current circulating in the winding, with the characteristic of maintaining this force approximately constant within the cursor work displacement range. The torque motor, with lower current and inductance values, has shorter response times than the servosolenoid, which operates with notably higher currents, but generates lower mechanical power outputs. The torque motor, therefore, constitutes the pilot stage usually found in servovalves, while the servosolenoid used in proportional valves acts directly on the valve spools.

The magnitude directly controlled by the servovalve or proportional valve can be a flow rate or a pressure difference, depending on the type.

Servovalves and proportional valves are usually distinguished on the basis of the following characteristics:

- input signals
- precision
- hysteresis
- linearity between input and output
- dead band
- bandwidth

Input signals are characterized by the type of signal and range of variation. Current signals (±10 mA or 4–20 mA) or voltage ones (0–10 V) are typical. Precision is intended as the difference between the desired value and the value effectively achieved. It is provided as a percentage of the full scale value. The hysteresis derives from the different behavior shown by the component with ascending settings and corresponding points descending. Its value expresses the percentage ratio between the maximum deviation and the full scale value. Linearity by nature is a characteristic that can be assessed over the entire working range. It can be expressed in an absolute manner as the maximum percentage deviation of the input/output relation of its linear regression. In general, better linearity is requested in position control compared with the cases of speed, pressure, or force controls. The dead band determines the minimum input value at which an output variation is obtained. Unlike the above, bandwidth is a

TABLE 20.11 Main Typical Differences Between Servovalves and Proportional Hydraulic Valves

	Servovalve	Proportional Valve
Electromechanical converter	Bidirectional torque motor (0.1 ÷ 0.2 W) with nozzle-flapper or jet pipe	Unidirectional servosolenoid (10 ÷ 20 W)
Input current	100 ÷ 200 mA	<3 A
Flow rate	2 ÷ 200 l/min (two stage type) with valve pressure drop = 70 bar	10 ÷ 500 l/min (single stage type) with valve pressure drop = 10 bar
Hysteresis	<3% (<1% with dither)	<6% (<2% with electric feedback)
Bandwidth	>100 Hz depending on the amplitude of the input and of the supply pressure	<100 Hz depending on the amplitude of the input
Radial clearance of the spool	1 μm (aerospace) 4 μm (industrial)	2 ÷ 6 μm
Dead band of the spool	<5% of the stroke	Overlap 10–20% of the stroke, less if compensated

characteristic of the dynamic type. This is because it refers to a frequency diagram of the component and defines the frequency at which the response drops by 3 dB below the low frequency value. Normally, a bandwidth two to five times greater is required for continuous control valves compared with that required by the system.

The main differences between servovalves and proportional valves are shown in Table 20.11. Except for the traditional difference in the electromechanical conversion device, there are overlaps in the static and dynamic characteristics in many components available on the market.

On the basis of the generically superior static and dynamic characteristics, servovalves are commonly used in closed loop controls while proportional valves are used in open loop systems.

Servovalves

Two-stage models are very common in the context of servovalves, where a first pilot stage converts a low power electric signal into a pressure difference capable of acting on the slide valve of the second stage, usually four-way and symmetrical. Flow rate control servovalves are divided into two categories on the basis of how they make the electric–hydraulic conversion:

- nozzle-flapper,
- jet-pipe.

An example of nozzle flapper servovalves is shown in Fig. 20.102. It comprises two stages: the former consists of a torque motor, the flapper, and a system of nozzles and chokes, while the latter consists mainly of the spool valve and output ports. The torque motor, constituted by the motor coil, the magnet, the armature, and the polepiece, is capable of transmitting a torque to the flapper which undergoes an angular displacement, thereby obstructing one of more calibrated nozzles to a greater degree. This operation causes a pressure difference at the ends of the spool, thereby causing the latter to move until the feedback wire, which connects the spool and the flapper, returns the flapper to the central position. Through the flow metering slot carried out in the bushing, the spool thereby permits communication between the various ports. The feedback wire is an elastic flexional element that provides the feedback between the main power stage (spool valve) and the first stage (torque motor).

This type of valve usually requires a greater degree of oil filtration, as nozzle-flappers are more sensitive to contaminants compared with the jet-pipe system.

Figure 20.103 shows a schematic section of an example of a servovalve of the jet-pipe type. It is connected to orifice P (supply) of the servovalve by means of a filter and flexible hose. It should be noted that, unlike the nozzle-flapper valve, it is not necessary to filter the entire incoming oil flow, but only what is called the control flow (the one going through the jet-pipe); this is certainly an advantage in terms of economic running and sensitivity to solid contamination. Starting with a standardized input voltage, the amplifier produces a voltage increase in one torque motor coil and an identical reduction in

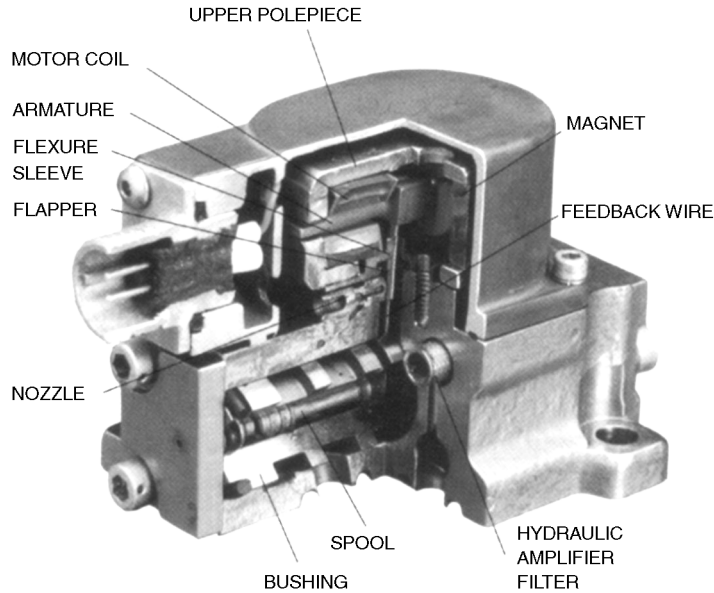



FIGURE 20.102 Nozzle-flapper servovalve (Moog).

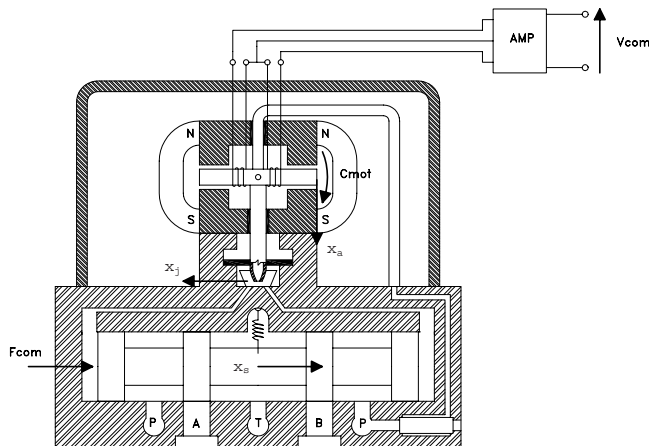



FIGURE 20.103 Jet-pipe servovalve scheme.

the other. This provokes an imbalance of the forces at the ends of the motor armature, generating a torque which tends to make the armature itself rotate and the jet-pipe with it. The displacement of the nozzle involves a different distribution of the control flow between the two pipes below it and, as a result, a pressure difference is created at the ends of the spool with a consequent net force on it, which causes its displacement. A spring element connects the spool and the jet-pipe, creating position feedback. The displacement of the spool and jet-pipe deforms the feedback spring, giving rise to a force proportional to it, and this makes it possible to balance the torque applied to the jet-pipe and the force generated at the ends of the spool. In this way, the system finds an equilibrium position, proportional to the input voltage. The feedback spring also produces centering of the slide valve at rest, making the presence of centering springs superfluous. Figure 20.104 shows the block diagram of the jet-pipe servovalve components with annotations of the physical magnitudes present in Fig. 20.103.

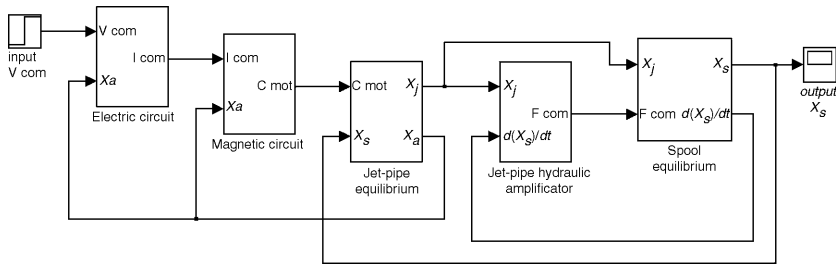



FIGURE 20.104 Block diagram of a jet-pipe servovalve.

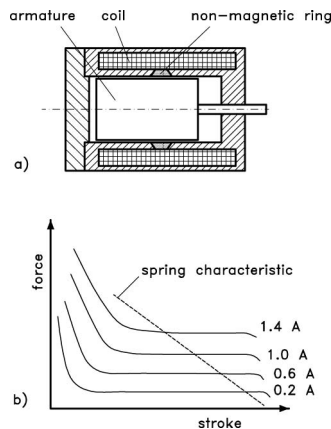



FIGURE 20.105 Proportional servosolenoid: (a) solenoid section scheme, (b) solenoid characteristics.

Proportional Valves

Proportional valves can be subdivided into proportional in flow and proportional in pressure (relief and pressure reduction) valves. In the first case, the action of the servosolenoid armature (Fig. 20.105(a)) displaces the main spool of the valve, which is checked by a spring on the valve body.

The characteristics of the force generated by the servosolenoid, along the entire possible stroke of the spool, is a function of the input current only, as indicated in Fig. 20.105(b). For all possible values of the input current, the equilibrium of the magnetic force supplied by the solenoid and of the feedback forces of the spring is determined by reaching a certain position value.

In cases in which precise positioning of the slide valve is requested, position feedback is introduced. The photograph of a proportional valve of this type is shown in Fig. 20.106. The input signal to the servosolenoid, sent by the feedback module, is the error compensated by a PID network between the reference signal and the feedback signal from the position transducer LVDT. The valve's accuracy and repeatability is improved by using the position feedback, as the hysteresis errors and those due to friction between the moving parts are partially compensated.

Flow proportional valves can have two stages. In this case, the outputs of the pilot proportional valve feed the end chambers of a spool valve of greater size, permitting greater controlled flows to be obtained while reducing dynamic performance at the same time. An example of a two-stage flow proportional valve is shown in Fig. 20.107.

In pressure regulator proportional valves, the action of the servosolenoid acts on a conical needle in such a way so as to regulate the pressure in the chamber upstream from the needle itself. Figure 20.108 shows the plan of a pilot operated proportional relief valve.

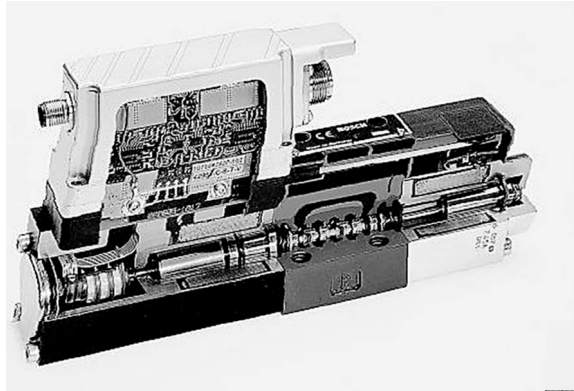



FIGURE 20.106 Flow proportional valve (Bosch Rexroth).

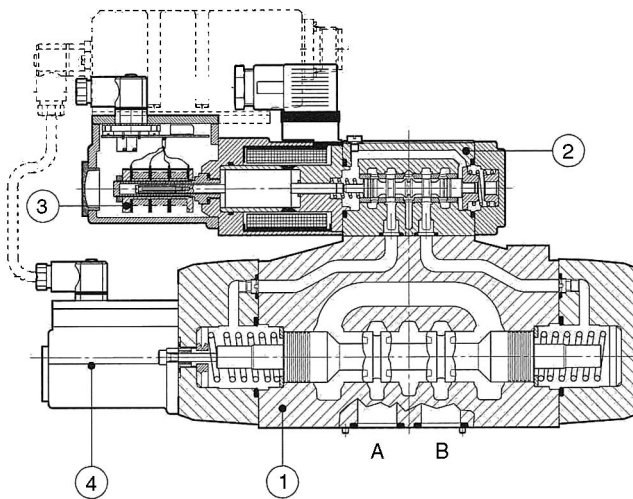



FIGURE 20.107 Double stage flow proportional valve: (1) main spool valve stage, (2) pilot stage, (3) LVDT of the pilot stage, (4) LVDT of the main stage (Atos).

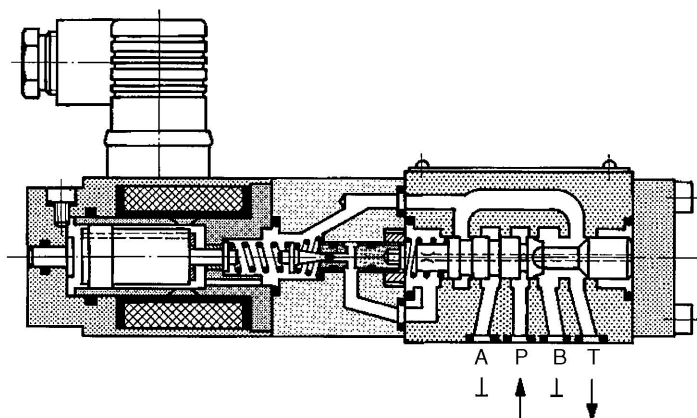



FIGURE 20.108 Proportional pressure relief valve (Bosch Rexroth).

Modeling of a Hydraulic Servosystem for Position Control

Figure 20.109 shows the scheme of a hydraulic servo system for position control constituted by a double acting, double ended actuator controlled in closed loop by a four-way servovalve. The x position of the piston is determined by the equilibrium of the forces acting on it: external force, thrust due to the pressures P_1 and P_2 acting in the chambers of the ram, friction force, and force of inertia. The pressures P_1 and P_2 are determined by the oil flows Q_{C1} and Q_{C2} entering and leaving the chambers. Flow rate Q_{FI} represents the oil leakage flow between the piston and the barrel. The flow proportional valve controls the oil flow on the basis of the reference signal *ref* from a compensator G_C. The input to the compensator is the error e_V between the signal V_{SET}, corresponding to the desired rod position x_{SET}, and the feedback signal $V_{F/B}$, corresponding to the effective rod position x measured by a position transducer LVDT.

The actuator is modeled by considering the equations of flow continuity in the chambers and the dynamic equilibrium equation for the rod. The continuity equation is expressed in a general form:

$$\rho\left(\sum Q_{IN} - \sum Q_{OUT}\right) = \frac{d(\rho V)}{dt} = \rho\frac{dV}{dt} + V\frac{d\rho}{dt} \tag{20.39}$$

where

$\sum Q_{IN}$ = sum of the flows in volume entering
$\sum Q_{OUT}$ = sum of the flows in volume leaving
ρ = density
V = volume
t = time

From the definition of the compressibility modulus of the oil β, P being the pressure in the chamber considered:

$$\frac{dV}{V} = -\frac{d\rho}{\rho} = -\frac{dP}{\beta} \tag{20.40}$$

We obtain

$$\sum Q_{IN} - \sum Q_{OUT} = \frac{dV}{dt} + \frac{V}{\beta}\frac{dP}{dt} \tag{20.41}$$

The continuity equation for chamber 1 is

$$Q_{C1} - Q_{FI} = A_c\dot{x} + \frac{V_0 + A_c x + V_{sm}}{\beta}\frac{dP_1}{dt} \tag{20.42}$$

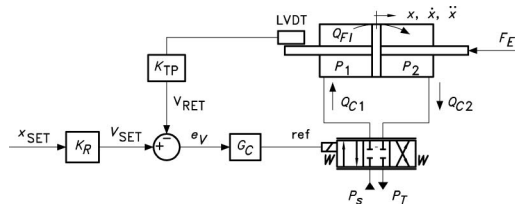



FIGURE 20.109 Scheme of a hydraulic servosystem with position control.

The continuity equation for chamber 2 is

$$Q_{FI} - Q_{C2} = A_c\dot{x} + \frac{V_0 - A_c x + V_{sm}}{\beta}\frac{dP_2}{dt} \tag{20.43}$$

where

Q_{C1} = flow entering chamber 1
Q_{C2} = flow leaving chamber 2
Q_{FI} = leakage flow between piston and barrel
A_c = thrust section = $(D_{al}^2 - D_{st}^2)\pi/4$
D_{al} = bore diameter
D_{st} = rod diameter
V_0 = volume of the chambers with piston centered = $A_c\,L/2$
L = stroke
x = piston displacement ($x = 0$ in centered position)
V_{sm} = dead band volume

The dynamic equilibrium equation of the piston is

$$P_1A_c - P_2A_c - F_e - M\ddot{x} - F_A = 0 \tag{20.44}$$

where

M = translating mass
F_e = external force
F_A = force of friction = $\gamma\dot{x} + F_{ATT}\,\mathrm{sign}(\dot{x})$
γ = coefficient of viscous friction
F_A = force of coulomb friction

Leaks can be modeled as resistances in laminar and steady-state conditions of the following type:

$$R = \frac{\Delta P}{Q} \tag{20.45}$$

where

R = resistance
Q = flow rate
ΔP = pressure difference

In the case of an annular pipe, we get

$$Q = \frac{\pi D h^3(1 + 1.5\varepsilon^2)}{12\mu l}\Delta P \tag{20.46}$$

where

D = seat diameter
h = meatus thickness = $(D - d)/2$
d = spool diameter
ε = eccentricity = $2e/(D - h)$
e = distance between seat axis and spool axis
μ = dynamic viscosity
l = meatus length

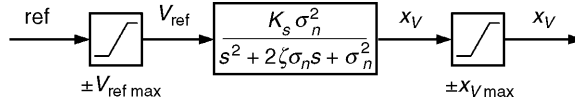



FIGURE 20.110 Block diagram of the valve.

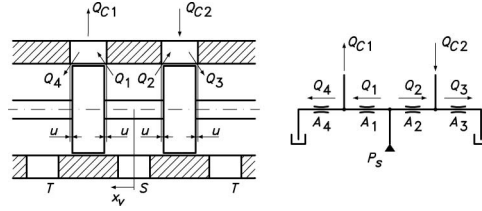



FIGURE 20.111 Reference scheme of the flow rates through the proportional valve.

The dynamic behavior of the electromechanical converter and of the spool valve can be identified with a linear model of the second order between the reference *ref* and the slide valve position x_V, as indicated in Fig. 20.110, where K_S (m/V) is the static gain of the valve, σ_n (rad/s) the natural frequency of the valve, and ζ the damping factor of the valve.

The flows regulated by the proportional valves are indicated in detail in the plan in Fig. 20.111. Having defined the areas A_1, A_2, A_3, A_4, functions of the spool displacement x_V, on the basis of the geometry, the flows transiting through the valve as a function of the pressures are given by

$$Q_1 = A_1 C_d \, \text{sign}(P_S - P_1) \sqrt{\frac{2}{\rho} |P_S - P_1|} \tag{20.47}$$

$$Q_4 = A_4 C_d \, \text{sign}(P_1 - P_T) \sqrt{\frac{2}{\rho} |P_1 - P_T|} \tag{20.48}$$

$$Q_2 = A_2 C_d \, \text{sign}(P_S - P_2) \sqrt{\frac{2}{\rho} |P_S - P_2|} \tag{20.49}$$

$$Q_3 = A_3 C_d \, \text{sign}(P_2 - P_T) \sqrt{\frac{2}{\rho} |P_2 - P_T|} \tag{20.50}$$

where P_S is the supply pressure, P_T is the pressure of the discharge reservoir, and C_d is the flow coefficient of the metering ports.

Therefore we get

$$Q_{C1} = Q_1 - Q_4 \tag{20.51}$$

$$Q_{C2} = Q_3 - Q_2 \tag{20.52}$$

The system of equations given above can be linearized in a working neighborhood, defined by the passage area of the valve A_{V0}, load pressure drop $P_{L0} = P_1 - P_2$, and piston position x_0. In the hypothesis

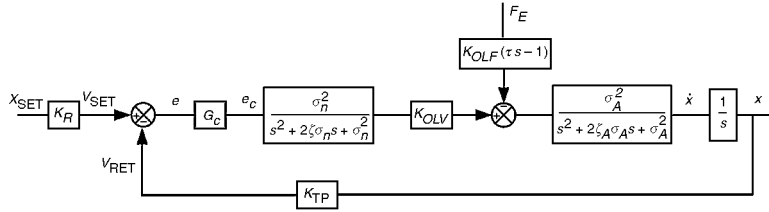



FIGURE 20.112 Block diagram of the linearized model of an hydraulic servosystem with position control.

that the leaks are negligible, the diagram of the linearized system is shown in Fig. 20.112, which indicates:

G_C	compensator transfer function [V/V]
$K_{OLV} = \dfrac{K'_S\, K_Q A_C}{A_C^2 - K_{PQ}\gamma}$	static speed gain $\left[\dfrac{\text{m/s}}{\text{V}}\right]$
$K_{TP} = K_R$	static gain of the position transducer [V/m]
$K_{OLF} = \dfrac{K_{PQ}}{A_C^2 - K_{PQ}\gamma}$	force constant $\left[\dfrac{\text{m}}{\text{sN}}\right]$
σ_n	natural frequency of the proportional valve [rad/s]
ζ	proportional valve damping factor
$\sigma_A = \sqrt{\dfrac{C_0}{M}[1-(K_{PQ}\gamma/A_C^2)]}$	hydraulic resonance frequency [rad/s]
$\zeta_A = \dfrac{\gamma-(MK_{PQ}C_0/A_C^2)}{2\sqrt{C_0 M[1-(K_{PQ}\gamma/A_C^2)]}}$	hydraulic damping factor
$\tau = \dfrac{A_C^2}{C_0 K_{PQ}}$	force disturbance time constant [s]

where, in the above expressions:

$K'_S = \dfrac{A_{V\max}}{V_{\max}}$	proportional valve static gain $\left[\dfrac{\text{m}^2}{\text{V}}\right]$
$K_Q = \left.\dfrac{\partial Q}{\partial A_V}\right\vert_{A_{V0},P_{L0}}$	proportional valve flow gain $\left[\dfrac{\text{m}^3}{\text{s m}^2}\right]$
A_C	piston area [m^2]
$K_{PQ} = \left.\dfrac{\partial Q}{\partial P_L}\right\vert_{A_{V0},P_{L0}}$	proportional valve flow-pressure gain $\left[\dfrac{\text{m}^3}{\text{s Pa}}\right]$
γ	coefficient of viscous friction $\left[\dfrac{\text{N}}{\text{s/m}}\right]$
$C_0 = \dfrac{4\beta A_C^2}{V_T}$	hydraulic stiffness in centred position [N/m]
V_T	total volume of the two chambers [m^3]
M	mass of the moving parts [kg]

The open loop transfer function is

$$G_{OL} = \frac{V_{RET}}{e} = G_C \cdot K_{OLV} \cdot K_{TP} \frac{\sigma_n^2}{s^2 + 2\zeta\sigma_n s + \sigma_n^2} \cdot \frac{\sigma_A^2}{s^2 + 2\zeta_A\sigma_A s + \sigma_A^2} \cdot \frac{1}{s} \tag{20.53}$$

Supposing G_C to be constant, the static gain in open loop is

$$K_0 = G_C \cdot K_{OLV} \cdot K_{TP} = \frac{G_C K_S' K_Q A_C K_{TP}}{A_C^2 - K_{PQ}\gamma} \tag{20.54}$$

The closed loop transfer function is

$$\begin{aligned} G_{CL} = \frac{x}{X_{\text{set}}} &= \frac{K_R G_C K_{OLV} \sigma_n^2 \sigma_A^2}{(s^2 + 2\zeta\sigma_n s + \sigma_n^2)(s^2 + 2\zeta_A\sigma_A s + \sigma_A^2)s + G_C K_{OLV}\sigma_n^2\sigma_A^2 K_{TP}} \\ &= \frac{1}{a_5 s^5 + a_4 s^4 + a_3 s^3 + a_2 s^2 + a_1 s + 1} \end{aligned} \tag{20.55}$$

where

$$a_5 = \frac{1}{\sigma_n^2\sigma_A^2 K_0}, \quad a_4 = \frac{2}{\sigma_n\sigma_A K_0}\left(\frac{\zeta_A}{\sigma_n} + \frac{\zeta}{\sigma_A}\right)$$

$$a_3 = \frac{1}{K_0}\left(\frac{1}{\sigma_n^2} + \frac{4\zeta_A\zeta}{\sigma_n\sigma_A} + \frac{1}{\sigma_A^2}\right), \quad a_2 = \frac{2}{K_0}\left(\frac{\zeta}{\sigma_n} + \frac{\zeta_A}{\sigma_A}\right), \quad a_1 = \frac{1}{K_0}$$

The transfer function between output and disturbance, said dynamic compliance, is

$$G_{FCL} = \frac{x}{F_e} = -\frac{K_{OLF}(\tau s - 1)(s^2 + 2\zeta\sigma_n s + \sigma_n^2)\sigma_A^2}{(s^2 + 2\zeta\sigma_n s + \sigma_n^2)(s^2 + 2\zeta_A\sigma_A s + \sigma_A^2)s + G_C K_{OLV}\sigma_n^2\sigma_A^2 K_{TP}} \tag{20.56}$$

Static compliance is

$$\left.\frac{x}{F_e}\right|_{s=0} = \frac{K_{OLF}}{G_C K_{OLV} K_{TP}} = \frac{K_{PQ}}{G_C K_S' K_Q A_C K_{TP}} = \frac{1}{G_C K_S' K_P A_C K_{TP}} \tag{20.57}$$

having put

$$K_{PQ} = \frac{K_Q}{K_P}, \quad K_P = \left.\frac{\partial P_L}{\partial A_V}\right|_{A_{V0},\, P_{L0}} \quad \text{pressure gain} \left[\frac{\text{Pa}}{\text{m}^2}\right]$$

Finally, a static stiffness is defined equal to

$$\left.\frac{F_e}{x}\right|_{s=0} = G_C K_S' K_P A_C K_{TP} \tag{20.58}$$

The predominant time constant, which is obtainable from the closed loop transfer function, is the coefficient $a_1 = 1/K_0$.

In conclusion, the following general considerations can be drawn:

- The speed gain K_{OLV}, and therefore the open loop static gain K_0, depend to a considerable degree on the flow gain K_Q and increase with increases in K_Q. K_Q increases as P_S increases, decreases as ΔP_{L0} increases, and does not vary with A_{V0}. In the hypothesis of γ below 1000 Ns/m, the effect of K_{PQ} is modest, practically negligible.
- The force constant K_{OLF} depends on the flow-pressure gain K_{PQ} and increases with it. $|K_{PQ}|$ increases with A_{V0} and with ΔP_{L0}, while it decreases as P_S increases; therefore $|K_{OLF}|$ decreases as P_S increases. Leaks lead to an increase in K_{OLF}.
- Static stiffness depends considerably on the pressure gain of the valve and increases with it. Given that K_P decreases with the leaks, these lead to a reduction in static stiffness. Furthermore, given that $|K_{PQ}|$ increases with A_{V0} while K_Q does not vary with A_{V0}, the pressure gain decreases with the increase in A_{V0}, and therefore static stiffness decreases if the valve is working in greater opening conditions.

  Furthermore, given that $|K_{PQ}|$ decreases as P_S increases, while K_Q increases as P_S increases, the pressure gain increases as P_S increases and therefore, the static stiffness increases with P_S.
- The hydraulic resonance frequency increases with the increase in hydraulic stiffness C_0 and decreases with the increase of the mass M. It is practically uninfluenced by the flow-pressure gain K_{PQ} if $\gamma < 1000$ Ns/m.
- The predominant time constant is inversely proportional to the speed gain, decreases as K_Q increases, that is it decreases as P_S increases, increases as ΔP_{L0} increases, and is indifferent to variations in A_{V0}.

Pneumatic Actuation Systems

Just as described for the hydraulic system, the components of a pneumatic actuation system are:

- the compressed air generation system, consisting of the compressor, the cooler, possibly a dryer, the storage tank, and the intake and output filters;
- the compressed air treatment unit, usually consisting of the FRL assembly (filter, pressure regulator, and possibly a lubrifier), which permits filtration and local regulation of the supply pressure to the actuator valve;
- the valve, that is, the regulator of the pneumatic power;
- the actuator, which converts the pneumatic power into mechanical power;
- the piping;
- the sensors and transducers;
- the system display, physical magnitude measurement, and control devices.

Some of the components of the pneumatic actuation system such as the compressors, treatment units, and some valves used in pneumatic servosystems are described below. The actuators are similar in function and construction to hydraulic ones, though they are built slightly lighter because of the lower working pressure.

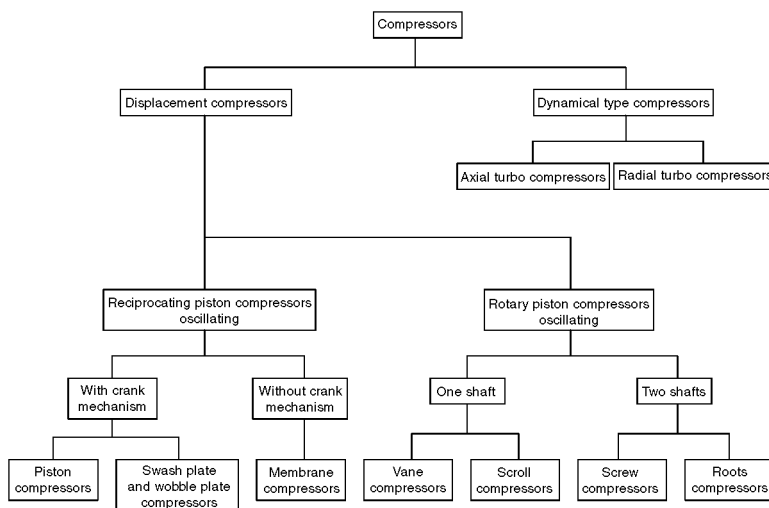



FIGURE 20.113 Classification of pneumatic compressors.

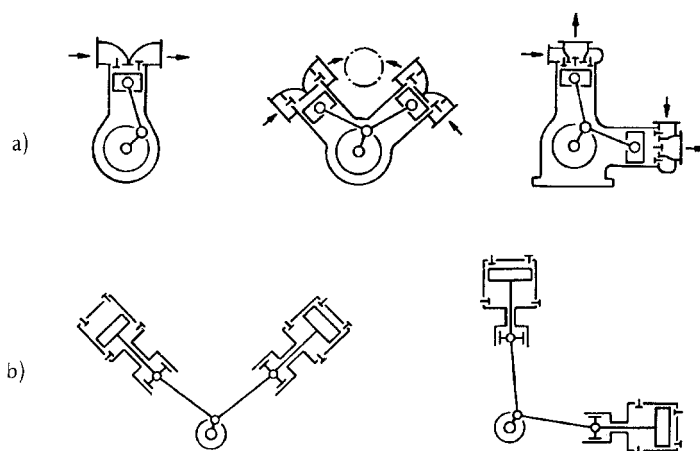



FIGURE 20.114 Piston compressors: (a) single action, (b) double action.

Compressors

The types of compressors used to produce compressed air are summarized in Fig. 20.113. In volumetric compressors, the air or gas is sucked in by means of a valve in the compression chamber where its volume is reduced to cause compression of the gas. Opening of the delivery valve, when a predetermined pressure has been reached, results in the distribution of the air mass to the user.

Vice versa, in dynamic compressors or turbocompressors, the kinetic energy is converted into pressure energy transferred to the gas as a result of the rotary motion of the impeller.

Alternating piston compressors determine the compression of the gas as an effect of the motion of the piston, moved by a connecting rod and crank mechanism, inside a gas-tight cylinder. They can be single and double acting, with one or more pistons and one or more stages (Fig. 20.114). They make it possible to obtain pressures of hundreds bar, where there are several stages, and flow rates of thousands of cubic meters per hour, in the case of several cylinders. Vane compressors (Fig. 20.115) have a rotor, fitted eccentrically with respect to the axis of the cylinder in which it rotates, which leads to a certain

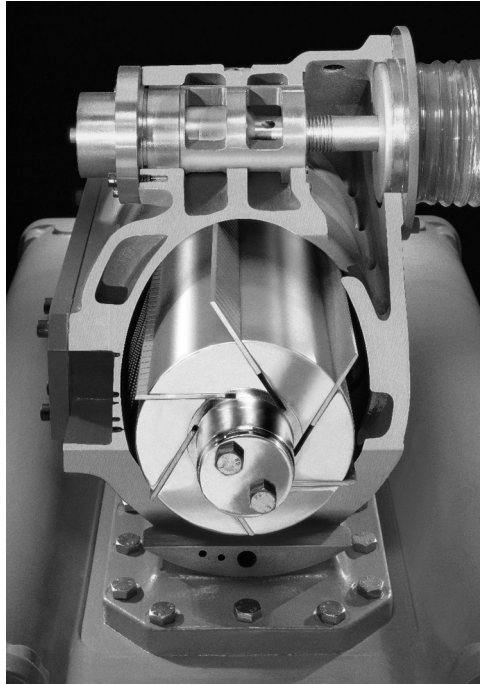



FIGURE 20.115 Rotary vane compressor (Pneumofore).

number of vanes which can move radially with respect to its axis. In the continuous rotation motion of the rotor, the vanes are centrifuged in contact with the seat of the stator, isolating chambers whose volume varies progressively with the angular stroke, guaranteeing input suction on the one hand, and a compressed gas output on the other. The compression pressures are below 15 bar, with maximum flow rates of 500 m^3/h. Compared with reciprocating piston compressors, they have less flow pulsation, fewer vibrations, and are more compact.

Screw compressors have two rotors rotating in opposite directions inside a stator, one with convex lobes and the other with concave lobes. The coupling of the profiles of the two rotors leads to a reduction of the volume during the angular stroke and consequent compression of the gas. With pressures typically below 15 bar, they provide a sufficiently continuous flow, up to values of about 3000 m^3/h.

In the same way, Roots compressors, also known as superchargers, are made up of two figure-of-eight-shaped rotors counter-rotating inside a stator in such a way as to transport volumes of gas from suction to delivery. Their efficiency is low because of the leakage between the rotors themselves and between the lobes and the casing, and they are, therefore, used for low compression pressures, below 2 bar. However, they do permit operation without lubrication, like screw compressors, so that oil-free air can be obtained.

Both axial and radial dynamic compressors are used to obtain high compressed air flow rates from a few thousand to 100,000 m^3/h.

Compressed Air Treatment Units

Pneumatic supply to a servosystem is generally provided by a local gas treatment unit, consisting of a filter, connected to a compressed gas distribution and generation network, a pressure regulator, and in case a lubricator *L*. Figure 20.116 shows an example of an integrated filter device and pressure regulator. The air first passes through the filter and is filtered by the deflector while the solid and liquid impurities in contact with the walls are deposited on the bottom of the cup, also as an effect of the conical bottom screen, located below the porous cylindrical element in sintered bronze or fabric. The filtered air then flows into the inlet of the pressure regulator, made up of an obturator in equilibrium between pressure

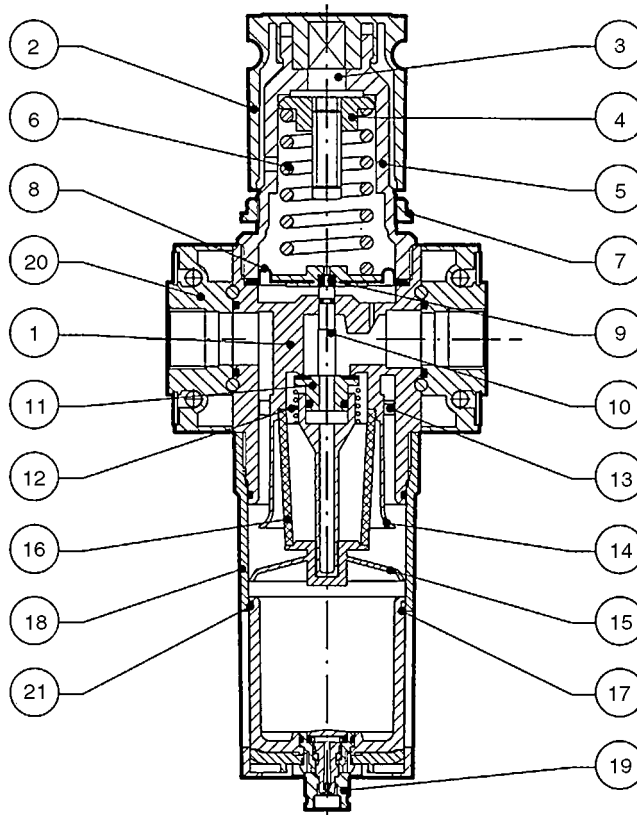



FIGURE 20.116 Pneumatic filter/pressure reducer (Metal Work).

forces. Control of the downstream pressure is determined by the position of the main obturator, which regulates the flow towards the outlet. The passage aperture is closed when the force due to the downstream pressure, acting on a diaphragm and on a translating piston, is in equilibrium with the force of the top spring, the preload of which is set by the rotation of the control knob. Vice versa, if the pressure force is below the desired value, the flow sent to the user tends to compensate for the pressure error with consequent closing of the obturator again when the set point has been reached. The opposite situation occurs if the regulated pressure is above the requested value, so that an aperture passage opens between the user and discharge.

Pneumatic Valves

Pneumatic valves are functionally similar to those used in hydraulic systems, so that reference should be made to the general considerations described above. In particular, this is also valid for the directional valves of the digital and proportional types. Even in pneumatic systems, there are digital spool or poppet two-, three-, or four-way distributors, with two or three working positions, and actuated manually, mechanically, pneumatically, and electrically.

Flow proportional valves are substantially similar to hydraulic ones and are available both with the torque motor electromechanical converter (servovalve), and with servosolenoid acting directly on the spool.

As well as these components for controlling the gas flow, digital electrically controlled two- or three-way valves are also used, and their control signals are modulated using PWM (pulse width modulation), PFM (pulse frequency modulation), PCM (pulse code modulation), PNM (pulse number modulation), or a combination of these.

As far as pressure regulation valves are concerned, three-way pressure proportional valves are available for pneumatic actuation which convert an electrical reference signal with standardized input into a controlled output pressure with good dynamics and high precision.

PWM (Pulse Width Modulation) Valves

The structure of PWM valves is similar to the corresponding electrically controlled unistable digital valves, but uses a technique for modulating the width of the pulses sent to the solenoid for supplying proportional control of the flow rate. This technique envisages that the input voltage reference analog signal V_{REF} (for example 0–10 V) is converted by a special driver into a digital V_{PWM} (ON/OFF) signal with pulse duration proportional to the input signal. Alternatively, the modulated signal can be generated directly by a digital controller, such as a PLC.

Figure 20.117 shows the PWM operating principle. The digital voltage signal sent to the valve solenoid is made up of a pulse train, with a constant amplitude, with a constant period T, but with the duration t of every pulse being a linear function of the analog value of the reference voltage. The average valve opening value, and therefore an initial approximation of the generated flow, is a function of the duration t of the pulse, in particular of the duty cycle t/T, and increases as the latter increases.

PWM valves generally do not have any feedback, so that the value of the downstream pressure, and therefore of the flow rate, depends on the type of pneumatic circuit present.

Figure 20.118 shows two plans which depict operation of the two-way, two-position valves, with PWM, used as a flow regulator (Fig. 20.118(a)) and as a pressure regulator (Fig. 20.118(b)). In plan a, the valve proportionally controls the flow which transits between the two points at pressure P_S (feed pressure) and at pressure P_V (downstream pressure) maintained constant. In this case, this flow is only a function of the aperture of the valve and therefore the proportionality is of the linear type. In plan b, the cross-fitted valves control the pressure P_R, for example, within a fluid capacity of volume V, respectively regulating the mass flow rate G_1 and discharge flow G_2. The time gradient for the controlled pressure P_R corresponds to the resulting flow G entering the reservoir.

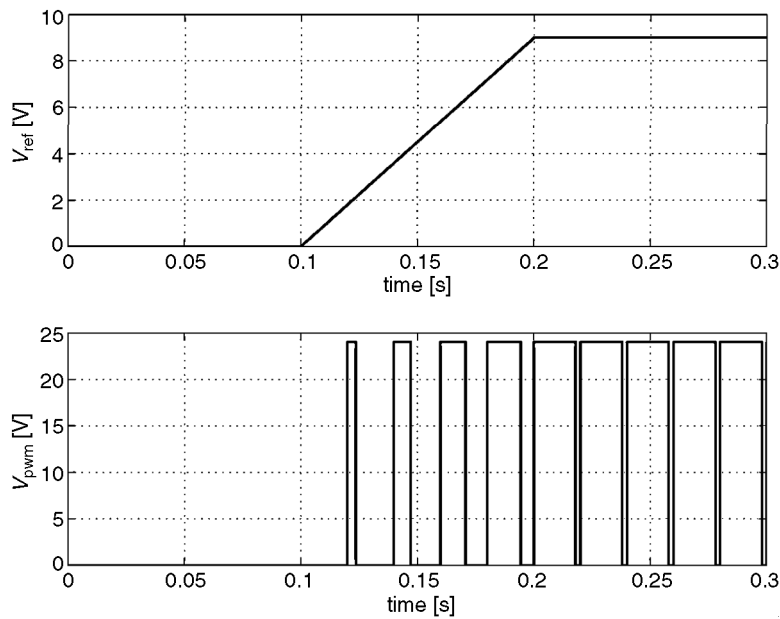



FIGURE 20.117 PWM (pulse width modulation) input and output signals.

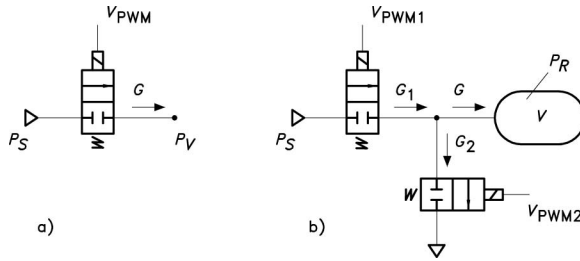



FIGURE 20.118 PWM (pulse width modulation) digital valves: (a) flow regulator, (b) pressure regulator.

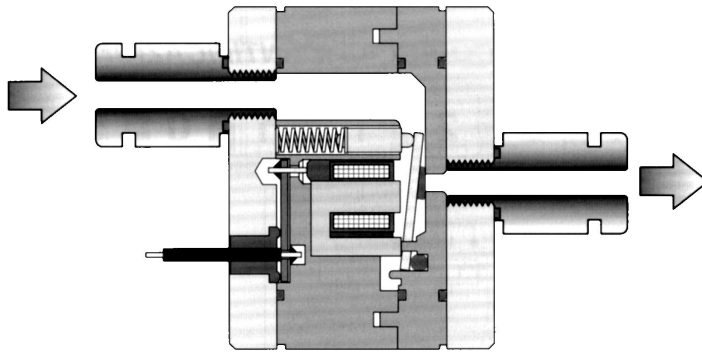



FIGURE 20.119 Two-way digital poppet valve (Matrix).

The parameters affecting the performance of a regulation made by a PWM on/off valve are as follows:

- valve opening and closing times
- dependence on the opening/closing times of the upstream and downstream pressures
- valve size
- period T or modulation carrier frequency $f = 1/T$
- working life of the valve

While small opening/closing times and high flow capacity are always antithetical characteristics in an on/off valve, when designing the system it is always necessary to find a compromise between the need for good control resolution and linearity and a high response dynamic.

In pneumatic servosystem applications, typical carrier frequency values $f = 1/T$ range between 20 and 100 Hz, so that valves with opening/closing times of 1 ÷ 5 ms are used.

An example of a normally closed 2/2 valve that guarantees minimum opening times (below 1 ms with speed-up command) can be seen in Fig. 20.119. These characteristics are obtained by reducing the mass of the moving parts, with the use of a poppet connected to a small oscillating bar, while practically eliminating the friction between the parts in relative motion.

Proportional Pressure Regulator Valves

These valves are normally three-way, with double poppets or with spool. Poppet valves operate in a similar way to pressure regulator valves. In the same way as with pressure regulators, the poppet which separates the high pressure environment from the regulated pressure one is in equilibrium between the force due to the regulated pressure and that exerted by the action of the control block. The latter can directly be the force of the servosolenoid armature, or that due to a pressure controlled by the control block which acts on a piston or on a diaphragm linked with the poppet.

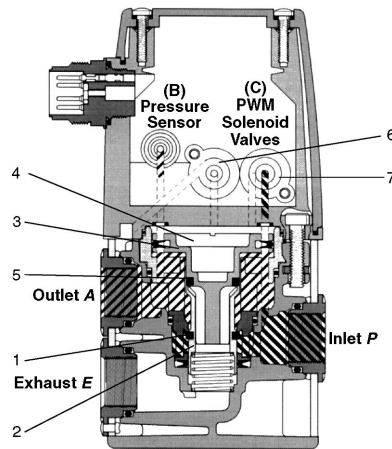



FIGURE 20.120 Pressure proportional valve (Parker).

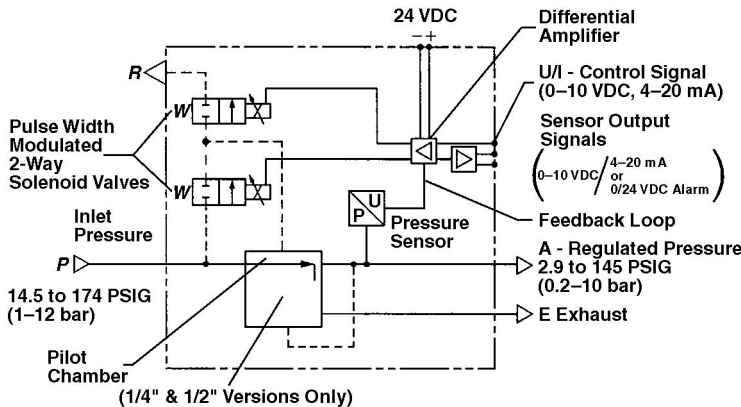



FIGURE 20.121 Pneumatic control scheme of the pressure proportional valve.

Figure 20.120 shows an example of a pressure proportional valve with double poppets. The ports at supply pressure, controlled pressure, and discharge are respectively indicated by P, A, and E. In the position indicated in the figure, the supply poppet 2 is at the top end of its stroke, as the seal 1 is against the fixed seat. In the same way, the regulating poppet 3 is in contact with the poppet 2 by means of the seal 5. The opening of the feed aperture, between the port P and the port A, is determined by the equilibrium of the forces acting on the piston of poppet 3, in particular the force F_R of the regulation pressure P_R in the servochamber 4 directed downwards, and the force F_C due to the action of the regulated pressure P_c on the outlet, directed upwards. If $F_R = F_C$, the moving bodies in the valve are in the positions shown in the figure, so that the chamber at controlled pressure P_C is isolated both from supply and discharge. If $F_R > F_C$, then the two poppets move downwards and the feed aperture is opened so as to convey the air mass towards the output and rebalance the pressure P_C at the desired value. In the opposite case, if $F_R < F_C$, the regulating poppet moves upwards, but while remaining at the top end of its stroke, the seal 5 opens and permits the passage of the masses from port A to the exhaust E.

In Fig. 20.120, 6 and 7 indicate the PWM on/off valves, which regulate the pressure P_R of the servochamber 4.

The pneumatic control plan of the valve is shown in Fig. 20.121. The two 2-way PWM valves receive the modulated control signal from the regulation block. These are fitted in such a way that one controls

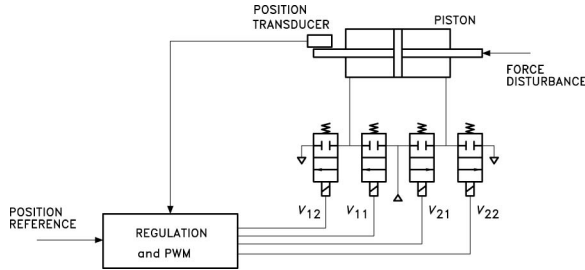



FIGURE 20.122 Scheme of a pneumatic servosystem with two-ways digital valves.

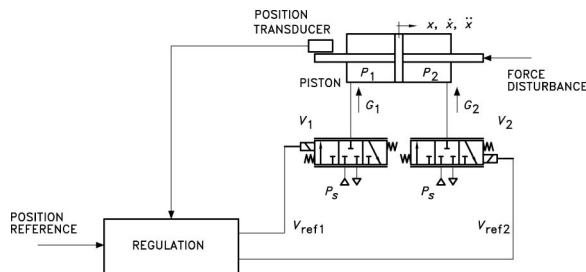



FIGURE 20.123 Scheme of a pneumatic servosystem with three-ways proportional valves.

a flow entering the servochamber 4 (see Fig. 20.120) while the other controls the flow exiting towards discharge. By means of appropriate action, the control signal is converted into a pressure proportional signal.

Modeling a Pneumatic Servosystem

The circuitry plan of a pneumatic servosystem capable of controlling the position, speed, or force can be similar to that shown in Fig. 20.109 for hydraulic actuation. The signal of the transducer of the desired magnitude must be specially fed back in a closed loop on the regulator depending on the controlled magnitudes.

In the plan in Fig. 20.122, the axial position of the piston, fed back by means of the position transducer, is determined by controlling the pressure in thrust chambers 1 and 2 by means of the flow proportional interfaces. The position reference is compared with the feedback signal and the error is compensated in a control regulator. On the basis of the valve opening strategy used, the signal is sent to the regulating valves which feed the chambers of the piston, hypothesized to be symmetrical. The pressure forces acting on the thrust surfaces of the piston oppose the external force disturbance. The circuitry plan hypothesizes the use of four digital valves each with two unistable ways, electrically controlled. This solution makes it possible to use small-sized valves, with resulting high bandwidth, which must be compatible with the overall bandwidth requested by the pneumatic servosystem. In this solution, the proportionality of the opening of the valves is obtained by pulse width modulation of the digital signal. Each pair of valves V_{11}, V_{12} and V_{21}, V_{22} constitutes a three-way valve the output of which is connected to a piston chamber, so that the scheme in Fig. 20.122 can be equivalent to that in Fig. 20.123 with the three-way analogically controlled valves V_1 and V_2.

The cylinder model envisages a system with three-differential equations, two of continuity of the air mass in the chambers and one of dynamic translation equilibrium.

The following magnitudes are differentiated (the subscripts 1 and 2 refer, respectively, to the rear 1 and front 2 chambers of the pistons):

A	piston thrust section
F_e	disturbance of force acting on the piston rod
G	mass flow rate of air entering the chamber
M	mass of the translating parts of the piston
n	air polytropic coefficient
P	cylinder chamber pressure
P_i	initial cylinder chamber pressure
P_{amb}	ambient pressure
R	air constant
T_i	initial cylinder chamber air temperature
x	rod position measured starting from x_0
x_0	piston half stroke
x_m	dead band
γ	coefficient of viscous friction

The continuity and equilibrium equations are given by:

$$\frac{dP_1}{dt} = \frac{G_1 nRT_{1i}}{A_1(x_0 + x_{m1} + x)(P_1/P_{1i})^{(1-n)/n}} - \frac{P_1 n}{(x_0 + x_{m1} + x)}\frac{dx}{dt} \tag{20.58}$$

$$\frac{dP_2}{dt} = \frac{G_2 nRT_{2i}}{A_2(x_0 + x_{m2} - x)(P_2/P_{2i})^{(1-n)/n}} + \frac{P_2 n}{(x_0 + x_{m2} - x)}\frac{dx}{dt} \tag{20.59}$$

$$\frac{d^2x}{dt^2} = \frac{(P_1 - P_{\text{amb}})A_1 - (P_2 - P_{\text{amb}})A_2 - F_e - \gamma dx/dt}{M} \tag{20.60}$$

The flow proportional valve V_1 is modeled as a variable section pneumatic resistance. The equations used for calculating mass flow rate G through a pneumatic resistance, characterized by a conductance C and by a critical ration b, in accordance with ISO 6358, which connects two environments A and B, with respective pressures of P_A and P_B, taken to be positive in the $A \rightarrow B$ direction, are

$$\text{sonic flow:} \quad G = \rho_0 P_A C \quad \text{for } 0 < \frac{P_B}{P_A} \leq b \tag{20.61}$$

$$\text{subsonic flow:} \quad G = \rho_0 P_A C\sqrt{1 - \left(\frac{P_B/P_A - b}{1 - b}\right)^2} \quad \text{for } b < \frac{P_B}{P_A} \leq 1 \tag{20.62}$$

$$\text{sonic flow:} \quad G = -\rho_0 P_B C \quad \text{for } 0 < \frac{P_A}{P_B} \leq b \tag{20.63}$$

$$\text{subsonic flow:} \quad G = -\rho_0 P_B C\sqrt{1 - \left(\frac{P_A/P_B - b}{1 - b}\right)^2} \quad \text{for } b < \frac{P_A}{P_B} \leq 1 \tag{20.64}$$

Hypothesizing a bipolar reference signal, it can be assumed that the range $V_{ref} > 0$ corresponds to the supply–user connection, while the field $V_{ref} < 0$ corresponds to the user–discharge connection. The appropriate equation of flow above must be rewritten in the same way.

Calculation of the conductance of the flow proportional valve is made considering the static and dynamic link between the reference voltage V_{ref} and the opening of the passage aperture A_V in accordance with modeling of the second order of the type:

$$\frac{d^2 A_V}{dt^2} + 2\zeta\sigma_n \frac{dA_V}{dt} + \sigma_n^2 A_V = K_s \sigma_n^2 V_{ref} \tag{20.65}$$

where ζ is the damping factor, σ_n is the valve's natural frequency, and K_s is its area static gain.

Assuming a static relation of the linear type between the opening A_V and the conductance C, as an initial approximation, we get

$$C = K_c A_V = K_c K_s V_{ref} = K_V V_{ref} \tag{20.66}$$

where K_V is the flow static gain of the valve, function of the maximum conductance C_{max}, and of the maximum value of the reference voltage $V_{ref\,max}$:

$$K_V = \frac{C_{max}}{V_{ref\,max}} \tag{20.67}$$

The complete dynamic relation between reference voltage and conductance is, therefore,

$$\frac{d^2 C}{dt^2} + 2\zeta\sigma_n \frac{dC}{dt} + \sigma_n^2 C = K_c \sigma_n^2 V_{ref} \tag{20.68}$$

The nonlinear model of the pneumatic servosystem with the position reference x_{set} and the force disturbance F_e as inputs, is made up of a nonlinear system of nine equations, of order eight overall, of the type:

a)	$C_1 = C_1(V_{ref\,1}, t)$	order 2	conductance of valve V_1 (see (20.68))
b)	$C_2 = C_2(V_{ref\,2}, t)$	order 2	flow rate of valve V_1 (see (20.68))
c)	$G_1 = G_1(C_1, P_1)$	order 0	flow rate of valve V_1 (see (20.61)–(20.64))
d)	$G_2 = G_2(C_2, P_2)$	order 0	flow rate of valve V_2 (see (20.61)–(20.64))
e)	$G_1 = G_1(P_1, \dot{P}_1, x, \dot{x})$	order 1	continuity chamber 1 (see (20.58))
f)	$G_2 = G_2(P_2, \dot{P}_2, x, \dot{x})$	order 1	continuity chamber 2 (see (20.59))
g)	$\ddot{x} = \ddot{x}(F_e, P_1, P_2, \dot{x})$	order 2	piston equilibrium (see (20.60))
h)	$V_{ref\,1} = V_{ref\,1}(x_{set}, x_{ret})$	order 0	V_1 valve control
i)	$V_{ref\,2} = V_{ref\,2}(x_{set}, x_{ret})$	order 0	V_2 valve control

If we want to carry out a linear analysis, it can be assumed that the equations *a*), *b*), *g*), *h*), *i*) are already written in linear form.

As far as the flow rates of valves *c*) and *d*) are concerned, it is hypothesized that the flow rate for each of them is subsonic in feed, with $V_{ref} > 0$, and sonic in discharge, with $V_{ref} < 0$. This means that for valve V_1, for example, the pressure P_1 must be within the range $bP_s < P_1 \le P_s$ in feed and in the range $P_1 \ge P_{amb}/b$

in discharge. This hypothesis is physically acceptable; hypothesizing $b = 0.3$, $P_s = 10$ bar, $P_{amb} = 1$ bar, we get that P_1 can vary between 3.33 bar and 10 bar. It is the same for P_2.

Linearizing the subsonic feed flow rate curve with a secant passing through the points $P_1 = P_s$, $G_1 = 0$ and $P_1 = b_1^* P_s$, $G_1 = G_{1\,\text{sonic}}$, of angular coefficient K_{L1}, we get

$$G_1 = K_{L1}(P_s - P_1) \tag{20.69}$$

where

$$K_{L1} = \frac{G_{1\,\text{sonic}}}{P_s - b_1^* P_s} = \frac{\rho_n C_1 P_s}{P_s(1 - b_1^*)} = \frac{\rho_n C_1}{1 - b_1^*} \tag{20.70}$$

or

$$G_1 = \frac{\rho_n (P_s - P_1)}{(1 - b_1^*)} C_1 \tag{20.71}$$

In the neighborhood $P_1 = P_{1r}$ and $C_1 = C_{1r} = 0$ (the subscript r indicates *of reference*), we get

$$G_1 = \frac{\rho_n (P_s - P_{1r})}{(1 - b_1^*)} C_1 = K_{11} C_1 \tag{20.72}$$

Equation (20.72) is valid for $V_{\text{ref1}} > 0$, to which $C_1 > 0$ corresponds analytically.

The flow rate discharge is expressed by

$$G_1 = \rho_n C_1 P_1 \tag{20.73}$$

which in the neighborhood $P_1 = P_{1r}$ becomes

$$G_1 = \rho_n P_{1r} C_1 = K_{12} C_1 \tag{20.74}$$

Equation (20.74) is valid for $V_{\text{ref1}} < 0$, to which $C_1 < 0$ corresponds analytically.

Calculating a mean of slopes K_{11} and K_{12} we get a mean slope K_{mean} given by

$$K_{\text{mean}} = \frac{K_{11} + K_{12}}{2} = \frac{\rho_n (P_s - b_1^* P_{1r})}{2(1 - b_1^*)} \tag{20.75}$$

The linearized flow rate as a function of C_1 therefore becomes

$$G_1 = \frac{\rho_n (P_s - b_1^* P_{1r})}{2(1 - b_1^*)} C_1 = K_1 C_1 \tag{20.76}$$

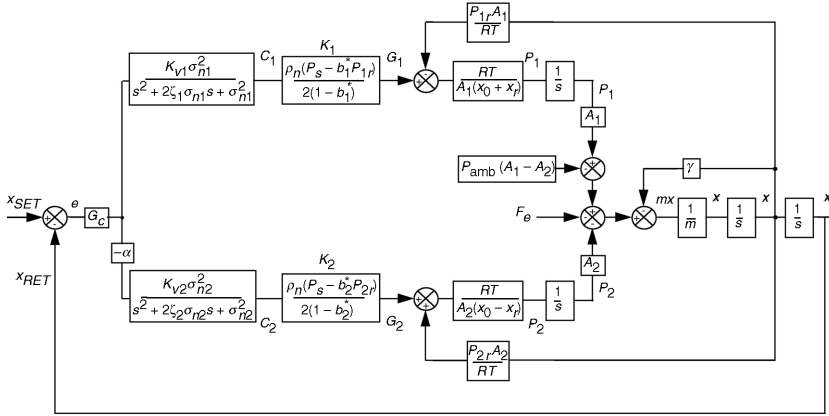



FIGURE 20.124 Block diagram of the linearized model of a pneumatic servosystem with position control.

In the same way for the valve V_2 we get

$$G_2 = \frac{\rho_n(P_s - b_2^* P_{2r})}{2(1 - b_2^*)} C_2 = K_2 C_2 \tag{20.77}$$

The continuity equations of the mass in the piston chambers e) and f), linearized in the reference neighborhood defined by

$$x = x_r, \quad P_1 = P_{1r}, \quad P_2 = P_{2r},$$

$$\dot{P}_1 = \dot{P}_{1r} = 0, \quad \dot{P}_2 = \dot{P}_{2r} = 0, \quad \dot{x} = \dot{x}_r = 0$$

$$x_{m1} = x_{m2} = 0, \quad n = 1$$

become

$$G_1 = \frac{P_{1r} A_1}{RT}\dot{x} + A_1 \frac{x_0 + x_r}{RT}\dot{P}_1 \tag{20.78}$$

$$G_2 = -\frac{P_{2r} A_2}{RT}\dot{x} + A_2 \frac{x_0 - x_r}{RT}\dot{P}_2 \tag{20.79}$$

The block diagram of the linearized model is shown in Fig. 20.124.

By applying the Laplace transforms of the system of linearized equations, assuming identical valves, we get:

$$\bar{x} = \frac{\sigma_n^2}{(s^2 + 2\zeta\sigma_n s + \sigma_n^2)} \frac{\sigma_A^2}{(s^2 + 2\zeta_A \sigma_A s + \sigma_A^2)} G_c K_{OLV}\, \bar{e} - \frac{\sigma_A^2}{(s^2 + 2\zeta_A \sigma_A s + \sigma_A^2)} K_{OLF} s \bar{F}_e + \text{C.I.} \tag{20.80}$$

where C.I. indicates the initial conditions, K_{OLV} is the static gain in speed, K_{OLF} is the gain of the force disturbance, σ_A and ζ_A are, respectively, the actuator's natural frequency and the damping factor, and G_c is the compensator law.

This result is shown in the block diagram in Fig. 20.125. Figure 20.126, on the other hand, shows the closed loop block diagram with position feedback. Obvious similarities can be seen when this plan is

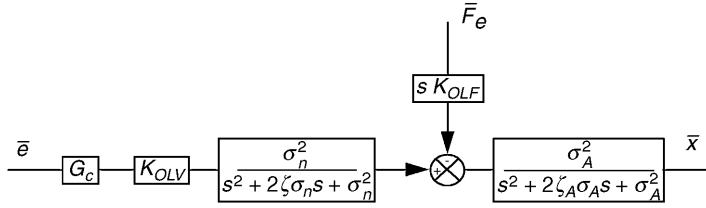



FIGURE 20.125 Block diagram of the open loop model.

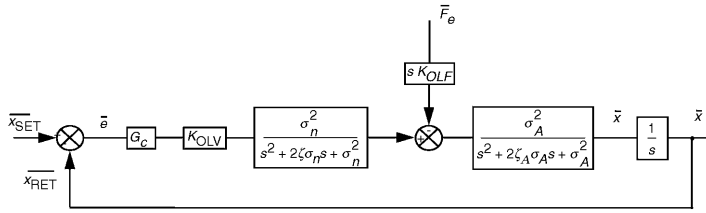



FIGURE 20.126 Block diagram of the closed loop model.

compared with that for a hydraulic servosystem.

In particular, hypothesizing that:

$P_{1r} = P_{2r} = \delta\, P_s$ with $\delta \in 0.6 - 0.9$
$A_1 = A_2 = A$ double ended actuator
$n = 1$ isothermal transformation

we can express the significant parameters of the pneumatic servosystem:

$$K_{OLV} = \frac{RT\rho_n}{2} K_V \frac{1}{(1-b^*)} \frac{1}{A} \frac{(1/\delta - b^*)}{(1-b^*)} \tag{20.81}$$

$$K_{OLF} = \frac{x_0[1-(x_r/x_0)^2]}{2\delta P_s A} \tag{20.82}$$

$$\sigma_A = \sqrt{\frac{2\delta P_s A}{x_0 m[1-(x_r/x_0)^2]}} \tag{20.83}$$

$$\zeta_A = \gamma \sqrt{\frac{x_0[1-(x_r/x_0)^2]}{8\delta P_s A m}} \tag{20.84}$$

On the basis of the design specifications, it is possible to choose the size and characteristics of the servosystem components, operating first on the linearized model, and then checking the complete effectiveness of the choice made with a complete nonlinear system model.

References

Andersen, B. W., *The analysis and design of pneumatic systems,* Wiley, New York, 1967.

Bouteille, D., Belforte, G., *Automazione flessibile, elettropneumatica e pneumatica,* Tecniche Nuove, Milano, 1987.

Belforte, G., D' Alfio, N., *Applicazioni e prove dell'automazione a fluido,* Levrotto & Bella, Torino, 1997.
Belforte, G., Manuello Bertetto, A., Mazza, L., *Pneumatica: corso completo,* Tecniche Nuove, Milano, 1998.
Blackburn, J. F., Reethof, G., Shearer, J. L., *Fluid power control,* MIT Press, Cambridge, 1960.
Dransfield, P., *Hydraulic control systems—design and analysis of their dynamics,* Springer, Berlin, 1981.
Esposito, A., *Fluid power with applications,* 5th ed., Prentice-Hall, Upper Saddle River, NJ, 2000.
Gotz, W., *Hydraulics. Theory and applications,* Robert Bosch Automation Technology Division Training, Ditzingen, 1998.
Hehn, A. H., *Fluid power troubleshooting,* 2nd ed., Dekker, New York, 1995.
Introduction to hydraulic circuits and components, The University of Bath, Bath, 2000.
Introduction to control for electrohydraulic systems, The University of Bath, Bath, 1999.
Jacazio, G., Piombo, B., *Meccanica applicata alle macchine 3: Regolazione e servomeccanismi,* Levrotto & Bella, Torino, 1994.
Johnson, J. E., *Electrohydraulic servo systems,* 2nd ed., Penton IPC, Cleveland, 1977.
Johnson, J. L., *Design of electrohydraulic systems for industrial motion control,* Penton IPC, Cleveland, 1991.
Johnson, J. L., *Basic electronics for hydraulic motion control,* Penton IPC, Cleveland, 1992.
Lewis, E. E., Stern, H., *Design of hydraulic control systems,* McGraw-Hill, New York, 1962.
Mang, T., Dresel, W., *Lubricants and lubrications,* Wiley-VCH, Weinheim, 2001.
McCloy, D., Martin, H. R., *The control of fluid power,* Longman, London, 1973.
Merritt, H. E., *Hydraulic control systems,* Wiley, New York, 1967.
Moog, *Technical Bulletins,* 101–152, Moog, New York.
Muller, R., *Pneumatics. Theory and applications,* Robert Bosch Automation Technology Division Training, Ditzingen, 1998.
Nervegna, N., *Oleodinamica e Pneumatica,* Politeko, Torino, 1999.
Parr, A., *Hydraulics and pneumatics: a technician's and engineer's guide,* 2nd ed., Butterworth Heinemann, Oxford, 1998.
Shetty, D., Kolk, R. A., *Mechatronics system design,* PWS publishing company, Boston, 1997.
Tonyan, M. J., *Electronically controlled proportional valves: selection and application,* Dekker, New York, 1985.
Viersma, T. J., *Analysis, synthesis and design of hydraulic servosystems and pipelines,* Elsevier, Amsterdam, 1980.
Yeaple, F., *Fluid power design handbook,* 3rd ed., Dekker, New York, 1996.

20.5 MEMS: Microtransducers Analysis, Design, and Fabrication*

Sergey Edward Lyshevski

Introduction

In many applications (from medicine and biotechnology to aerospace and security), the use of nano- and microscale structures, devices, and systems is very important [1–4]. This chapter discusses the analysis, modeling, design, and fabrication of electromagnetic-based microscale structures and devices (microtransducers controlled by ICs). It is obvious that to attain our objectives and goals, the synergy of multidisciplinary engineering, science, and technology must be utilized. In particular, electromagnetic

*This section is a part of the book: S. E. Lyshevski, *MEMS and NEMS: Systems, Devices, and Structures,* CRC Press, Boca Raton, FL, 2001.

theory and mechanics comprise the fundamentals for analysis, modeling, simulation, design, and optimization, while fabrication is based on the micromachining and high-aspect-ratio techniques and processes, which are the extension of the CMOS technologies developed to fabricate ICs. For many years, the developments in microelectromechanical systems (MEMS) have been concentrated on the fabrication of microstructures adopting, modifying, and redesigning silicon-based processes and technologies commonly used in integrated microelectronics. The reason for refining of conventional processes and technologies as well as application of new materials is simple: in general, microstructures are three-dimensional with high aspect ratios and large structural heights in contrast to two-dimensional planar microelectronic devices. Silicon structures can be formed from bulk silicon micromachining using wet or dry processes, or through surface micromachining. Metallic micromolding techniques, based upon photolithographic processes, are also widely used to fabricate microstructures. Molds are created in polymer films (usually photoresist) on planar surfaces, and then filled by electrodepositing metal (electrodeposition plays a key role in the fabrication of the microstructures and microdevices, which are the components of MEMS). High-aspect ratio technologies use optical, e-beam, and x-ray lithography to create trenches up to 1 mm deep in polymethylmethacrylate resist on the electroplating base (called seed layer). Electrodeposition of magnetic materials and conductors, electroplating, electroetching, and lift-off are extremely important processes to fabricate microscale structures and devices. Though it is recognized that the ability to use and refine existing microelectronics fabrication technologies and materials is very important, and the development of novel processes to fabricate MEMS is a key factor in the rapid growth of affordable MEMS, other emerging areas arise. In particular, devising, design, modeling, analysis, and optimization of novel MEMS are extremely important. Therefore, recently, the MEMS theory and microengineering fundamentals have been expanded to thoroughly study other critical problems such as the system-level synthesis and integration, synergetic classification and analysis, modeling and design, as well as optimization. This chapter studies the fabrication, analysis, and design problems for electromagnetic microstructures and microdevices (microtransducers with ICs). The descriptions of the fabrication processes are given, modeling and analysis issues are emphasized, and the design is performed.

Design and Fabrication

In MEMS, the fabrication of thin film magnetic components and microstructures requires deposition of conductors, insulators, and magnetic materials. Some available bulk material constants (conductivity σ, resistivity ρ at 20°C, relative permeability μ_r, thermal expansion t_e, and dielectric constant—relative permittivity ε_r) in SI units are given in Table 20.12.

TABLE 20.12 Material Constants

Material	σ	ρ	μ_r	$t_e \times 10^{-6}$	ε_r
Silver	6.17×10^7	0.162×10^{-7}	0.9999998	NA	
Copper	5.8×10^7	0.172×10^{-7}	0.99999	16.7	
Gold	4.1×10^7	0.244×10^{-7}	0.99999	14	
Aluminum	3.82×10^7	0.26×10^{-7}	1.00000065	25	
Tungsten	1.82×10^7	0.55×10^{-7}	NA	NA	
Zinc	1.67×10^7	0.6×10^{-7}	NA	NA	
Cobalt	NA	NA	250	NA	
Nickel	1.45×10^7	0.69×10^{-7}	600 nonlinear	NA	
Iron	1.03×10^7	1×10^{-7}	4000 nonlinear	NA	
Si				2.65	11.8
SiO_2				0.51	3.8
Si_3N_4				2.7	7.6
SiC				3.0	6.5
GaAs				6.9	13
Ge				2.2	16.1

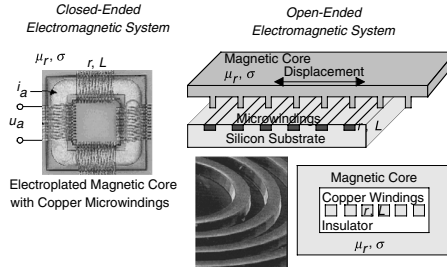



FIGURE 20.127 Closed-ended and open-ended electromagnetic systems in microtransducers (toroidal microstructures with the insulated copper circular conductors wound around the magnetic material and linear micromotor) with magnetic cores (stator and rotor electroplated thin films).

Although MEMS topologies and configurations vary (see the MEMS classification concept [2]), in general, electromagnetic microtransducers have been designed as the closed-ended, open-ended, and integrated electromagnetic systems. As an example, Fig. 20.127 illustrates the microtoroid and the linear micromotor with the closed-ended and open-ended electromagnetic systems, respectively. The copper windings and magnetic core (microstructures) can be made through electroplating, and Fig. 20.129 depicts the electroplated circular copper conductors which form the windings (10 μm wide and thick with 10 μm spacing) deposited on the insulated layer of the magnetic core.

The comprehensive electromagnetic analysis must be performed for microscale structures and devices. For example, the torque (force) developed and the voltage induced by microtransducers depend upon the inductance, and the microdevice's efficiency is a function of the winding resistance, resistivity of the coils deposited, eddy currents, hysteresis, etc. Studying the microtoroid, consider a circular path of radius R in a plane normal to the axis. The magnetic flux intensity is calculated using the following formula:

$$\oint_s \mathbf{H} \cdot d\mathbf{s} = 2\pi RH = Ni$$

where N is the number of turns. Thus, one has

$$H = \frac{Ni}{2\pi R}$$

The value of H is a function of R, and therefore, the field is not uniform.

Microwindings must guarantee the adequate inductance in the limited footprint area with the minimal resistance. For example, in the microtransducers and low power converters, 0.5 μH (or higher) inductance is required at high frequency (1–10 MHz). Compared with the conventional minidevices, the thin film electromagnetic microtransducers have lower efficiency due to higher resistivity of thin films, eddy currents, hysteresis, fringing effect, and other undesirable phenomena, which usually have the secondary (negligible) effect in the miniscale and conventional electromechanical devices. The inductance can be increased by ensuring a large number of turns, using core magnetic materials with high relative permeability, increasing the cross-sectional core area, and decreasing the path length. In fact, at low frequency, the formula for inductance is

$$L = \frac{\mu_0 \mu_r N^2 A}{l}$$

where μ_r is the relative permeability of the core material, A is the cross-sectional area of the magnetic core, and l is the magnetic path length.

Using the reluctance $\Re = l/(\mu_0 \mu_r A)$, one has $L = N^2/\Re$. For the electromagnetic microtransducers, the flux is a very important variable, and using the *net* current, one has $\Phi = Ni/\Re$.

It is important to recall that the inductance is related to the energy stored in the magnetic field, and

$$L = \frac{2W_m}{i^2} = \frac{1}{i^2}\int_v \mathbf{B}\cdot\mathbf{H}\,dv.$$

Thus, one has

$$L = \frac{1}{i^2}\int_v \mathbf{B}\cdot\mathbf{H}\,dv = \frac{1}{i^2}\int_v \mathbf{H}\cdot(\nabla\times\mathbf{A})\,dv = \frac{1}{i^2}\int_v \mathbf{A}\cdot\mathbf{J}\,dv = \frac{1}{i}\oint_l \mathbf{A}\cdot d\mathbf{l} = \frac{1}{i}\oint_s \mathbf{B}\cdot d\mathbf{s} = \frac{\Phi}{i}$$

or

$$L = \frac{N\Phi}{i}$$

We found that the inductance is the function of the number of turns, flux, and current.

Making use of the equation

$$L = \frac{\mu_0\mu_r N^2 A}{l}$$

one concludes that the inductance increases as a function of the squared number of turns. However, a large number of turns requires the high turn density (small track width and spacing so that many turns can be fitted in a given footprint area). However, reducing the track width leads to an increase in the conductor resistance, decreasing the efficiency. Therefore, the design trade-off between inductance and winding resistance must be studied. To achieve low resistance, one must deposit thick conductors with the thickness in the order of tens of micrometers. In fact, the dc resistance is found as $R = \rho_c\, l_c/A_c$, where ρ_c is the conductor resistivity, l_c is the conductor length, A_c is the conductor cross-sectional area. Therefore, the most feasible process for deposition of conductors is electroplating. High-aspect-ratio processes ensure thick conductors and small track widths and spaces (high-aspect-ratio conductors have a high thickness to width ratio). However, the footprint area is limited not allowing to achieve a large conductor cross-sectional area. High inductance value can also be achieved by increasing the magnetic core cross-sectional area using thick magnetic cores with large A. However, most thin film magnetic materials are thin film metal alloys, which generally have characteristics not as good as the bulk ferromagnetic materials. This results in the eddy current and undesirable hysteresis effects, which increase the core losses and decrease the inductance. It should be emphasized that eddy currents must be minimized.

As illustrated, magnetic cores and microwindings are key components of microstructures, and different magnetic and conductor materials and processes to fabricate microtransducers are employed. Commonly, the *permalloy* ($\text{nickel}_{80\%}$-$\text{iron}_{20\%}$ alloy) thin films are used. It should be emphasized that *permalloy* as well as other materials (e.g., amorphous cobalt-phosphorous) are soft magnetic materials that can be made through electrodeposition. In general, the deposits have nonuniform thickness and composition due to the electric current nonuniformity over the electrodeposition area. Furthermore, hydrodynamic effects in the electrolyte also usually increase nonuniformity (these nonuniformities are reduced by choosing a particular electrochemicals). The inductance and losses remain constant up to a certain frequency (which is a function of the layer thickness, materials used, fabrication processes, etc.), and in the high frequency operating regimes, the inductance rapidly decreases and the losses increase due to the eddy current and hysteresis effects. For example, for the *permalloy* ($Ni_{80\%}Fe_{20\%}$) thin film magnetic core and copper winding,

the inductance decreases rapidly above 1, 3, and 6 MHz for the 10, 8, and 5 μm thick layers, respectively. It should be emphasized that the skin depth of the magnetic core thin film as a function of the magnetic properties and the frequency f is found as $\delta = \sqrt{1/(\pi f \mu \sigma)}$, where μ and σ are the permeability and conductivity of the magnetic core material, respectively. The total power losses are found using the Pointing vector $\mathbf{\Xi} = \mathbf{E} \times \mathbf{H}$, and the total power loss can be approximately derived using the expression for the power crossing the conductor surface within the area, e.g.,

$$P_{\text{average}} = \int_s \Xi_{\text{average}} ds = \frac{1}{4}\int_s \sigma \delta E_o^2 e^{-2/\delta \sqrt{\pi f \mu \sigma}} ds$$

It is important to emphasize that the skin depth (depth of penetration) is available, and for the bulk copper $\delta_{\text{cu}} = 0.066/\sqrt{f}$.

In general, the inductance begins to decrease when the ratio of the lamination thickness to skin depth is greater than one. Thus, the lamination thickness must be less than skin depth at the operating frequency f to attain the high inductance value. In order to illustrate the need to comprehensively study microinductors, we analyze the toroidal microinductor (1 mm by 1 mm, 3 μm core thickness, 2000 permeability). The inductance and winding resistances are analyzed as the functions of the operating frequency. Modeling results indicate that the inductance remains constant up to 100 kHz and decreases for the higher frequency. The resistance increases significantly at frequencies higher than 150 kHz (the copper microconductor thickness is 2 μm, and the dc winding resistance is 10 Ω). The decreased inductance and increased resistance at high frequency are due to hysteresis and eddy current effects.

The skin depth in the magnetic core material depends on the permeability and the conductivity. The $Ni_{x\%}Fe_{100-x\%}$ thin films have a relative permeability in the range from 600 to 2000, and the resistivity is in the order of 20 $\mu\Omega$ cm. It should be emphasized that the materials with high resistivity have low eddy current losses and allow one to deposit thicker layers as the skin depth is high. Therefore, high resistivity magnetic materials are under consideration, and the electroplated FeCo thin films have 100–130 $\mu\Omega$ cm resistivity. Other high resistivity materials, which can be deposited by sputtering, are FeZrO and CoHfTaPd (sputtering has advantages for the deposition of laminated layers of magnetic and insulating materials because magnetic and insulating materials can be deposited in the same process step). Electroplating, as a technique for deposition of laminated multilayer structures, in general, requires different processes to deposit magnetic and insulating materials (layers).

The major processes involved in the electromagnetic microtransducer fabrication are etching and electroplating magnetic vias and through-holes, and then fabricating the inductor-type microstructures on top of the through-hole wafer using multilayer thick photoresist processes [5–7]. For example, let us use the silicon substrate (100-oriented *n*-type double-sided polished silicon wafers) with a thin layer of thermally grown silicon dioxide (SiO_2). Through-holes are patterned on the topside of the $Si–SiO_2$ wafer (photolithography process) and then etched in the KOH system (different etch rate can be attained based upon the concentration and temperature). Then, the wafer is removed from the KOH solution with 20–30 μm of silicon remaining to be etched. A seed layer of Ti–Cu or Cr–Cu (20–40 nm and 400–500 nm thickness, respectively) are deposited on the backside of the wafer using electron beam evaporation. The copper acts as the electroplating seed layer, while a titanium (or chromium) layer is used to increase adhesion of the copper layer to the silicon wafer. On the copper seed layer, a protective NiFe thin film layer is electroplated directly above the through-holes to attain protection and stability. The through-holes are fully etched again (in the KOH system), and then the remaining SiO_2 is stripped (using the BHF solution) to reveal the backside metal layers. Then, the titanium adhesion layer is etched in the HF solution (chromium, if the Cr–Cu seed layer is used, can be removed using the $K_3Fe(CN)_6$–NaOH solution). This allows the electroplating of through-holes from the exposed copper seed layer. The empty through-holes are electroplated with NiFe thin film. This forms the magnetic vias. Because the KOH-based etching process is crystallographically dependent, the sidewalls of the electroplating mold are the 111-oriented crystal planes (54.7° angular orientation to the surface). As a result of these 54.7°-angularly

oriented sidewalls, the electroplating can be nonuniform. To overcome this problem, the through-holes can be over-plated and polished to the surface level [5–7]. After the through-hole plating and polishing, the seed layer is removed, and 10–20 μm coat (e.g., polyimide PI2611) is spun on the backside and cured at 300°C to cover the protective NiFe layer. Now, the microinductor can be fabricated on the topside of the wafer. In particular, the microcoils are fabricated on top of the through-hole wafer with the specified magnetic core geometry (e.g., plate- or horseshoe-shaped) parallel to the surface of the wafer. The microcoils must be wounded around the magnetic core to form the electromagnetic system. Therefore, the additional structural layers are needed (for example, the first level is the conductors that are the bottom segments of each microcoil turn, the second level includes the magnetic core and vertical conductors which connect the top and bottom of each microcoil turn segment, and the third level consists of the top conductors that are connected to the electrical vias, and thus form microcoil turns wounded around the magnetic core). It is obvious that the insulation (dielectric) layers are required to insulate the magnetic core and microcoils. The fabrication can be performed through the electron beam evaporation of the Ti-Cu seed layer, and then, 25–35 μm electroplating molds are formed (AZ-4000 photoresist can be used). The copper microcoils are electroplated on the top of the mold through electroplating. After electroplating is completed, the photoresist is removed with acetone. Then, the seed layer is removed (copper is etched in the H_2SO_4 solution, while the titanium adhesion layer is etched by the HF solution). A new layer of the AZ-4000 photoresist is spun on the wafer to insulate the bottom conductors from the magnetic core. The vias' openings are patterned at the ends of the conductors, and the photoresist is cured forming the insulation layer. In addition to insulation, the hard curing leads to reflow of the photoresist serving the planarization purpose needed to pattern additional layers. Another seed layer is deposited from which electrical vias and magnetic core are patterned and electroplated. This leads to two lithography sequential steps, and the electrical vias (electroplated Cu) and magnetic core (NiFe thin film) are electroplated using the same seed layer. After the vias and magnetic core are completed, the photoresist and seed layers are removed. Then, the hard curing is performed. The top microconductors are patterned and deposited from another seed layer using the same process as explained above for the bottom microconductors. The detailed description of the processes described and the fabricated microtransducers are available in [5–7]. We have outlined the fabrication of microinductors because these techniques can be adopted and used to fabricate microtransducers. It also must be emphasized that the analysis and design can be performed using the equations given.

Analysis of Translational Microtransducers

Figure 20.128 illustrates a microelectromechanical device (translational microtransducer) with a stationary member (magnetic core with windings) and movable member (microplunger), which can be fabricated using the micromachining technology. Our goal is to perform the analysis and modeling of the microtransducer developing the lamped-parameter mathematical model. That is, the goal is to derive the differential equations which model the microtransducer steady-state and dynamic behavior.

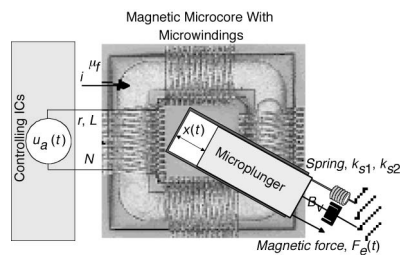



FIGURE 20.128 Schematic of the translational microtransducer with controlling ICs.

Applying Newton's second law for translational motion, we have

$$F(t) = m\frac{d^2x}{dt^2} + B_v\frac{dx}{dt} + (k_{s1}x + k_{s2}x^2) + F_e(t)$$

where x denotes the microplunger displacement, m is the mass of a movable member (microplunger), B_v is the viscous friction coefficient, k_{s1} and k_{s2} are the spring constants, and $F_e(t)$ is the magnetic force

$$F_e(i,x) = \frac{\partial W_c(i,x)}{\partial x}$$

It should be emphasized that the restoring/stretching force exerted by the spring is given by $(k_{s1}x + k_{s2}x^2)$.

Assuming that the magnetic system is linear, the coenergy is found to be $W_c(i, x) = \frac{1}{2}L(x)i^2$ and the electromagnetic force developed is given by

$$F_e(i,x) = \frac{1}{2}i^2\frac{dL(x)}{dx}$$

In this formula, the analytic expression for the term $dL(x)/dx$ must be found. The inductance is

$$L(x) = \frac{N^2}{\Re_f + \Re_g} = \frac{N^2\mu_f\mu_0A_fA_g}{A_gl_f + 2A_f\mu_f(x+2d)}$$

where $\Re_f$ and $\Re_g$ are the reluctances of the magnetic material and air gap; A_f and A_g are the cross-sectional areas; l_f and $(x + 2d)$ are the lengths of the magnetic material and the air gap.

Thus

$$\frac{dL}{dx} = -\frac{2N^2\mu_f^2\mu_0A_f^2A_g}{[A_gl_f + 2A_f\mu_f(x+2d)]^2}$$

Using Kirchhoff's law, the voltage equation for the electric circuit is

$$u_a = ri + \frac{d\psi}{dt}$$

where the flux linkage ψ is $\psi = L(x)i$.

Thus, one obtains

$$u_a = ri + L(x)\frac{di}{dt} + i\frac{dL(x)}{dx}\frac{dx}{dt}$$

Therefore, the following nonlinear differential equation results:

$$\frac{di}{dt} = -\frac{r}{L(x)}i + \frac{2N^2\mu_f^2\mu_0A_f^2A_g}{L(x)[A_gl_f + 2A_f\mu_f(x+2d)]^2}iv + \frac{1}{L(x)}\mu_a$$

Augmenting this equation with the differential equation and the *torsional-mechanical* dynamics

$$F(t) = m\frac{d^2x}{dt^2} + B_v\frac{dx}{dt} + (k_{s1}x + k_{s2}x^2) + F_e(t)$$

three nonlinear differential equations for the considered translational microtransducer are found to be

$$\frac{di}{dt} = -\frac{r[A_g l_f + 2A_f\mu_f(x+2d)]}{N^2\mu_f\mu_0 A_f A_g}i + \frac{2\mu_f A_f}{A_g l_f + 2A_f\mu_f(x+2d)}iv + \frac{A_g l_f + 2A_f\mu_f(x+2d)}{N^2\mu_f\mu_0 A_f A_g}u_a$$

$$\frac{dv}{dt} = \frac{N^2\mu_f^2\mu_0 A_f^2 A_g}{m[A_g l_f + 2A_f\mu_f(x+2d)]^2}i^2 - \frac{1}{m}(k_{s1}x + k_{s2}x^2) - \frac{B_v}{m}v$$

$$\frac{dx}{dt} = v$$

The derived differential equations represent the lumped-parameter mathematical model of the microtransducer. Although, in general, the high-fidelity modeling must be performed integrating nonlinearities (for example, nonlinear magnetic characteristics and hysteresis) and secondary effects, the lumped-parameter mathematical models as given in the form of nonlinear differential equations have been validated for microtransducers. It is found that the major phenomena and effects are modeled for the current, velocity, and displacement (secondary effects such as Coulomb friction, hysteresis and eddy currents, fringing effect and other phenomena have not been modeled and analyzed). However, the lumped-parameter modeling provides one with the capabilities to attain reliable preliminary steady-state and dynamic analysis using primary circuitry and mechanical variables. It is also important to emphasize that the voltage, applied to the microwinding, is regulated by ICs. The majority of ICs to control microtransducers are designed using the pulse-width-modulation topologies. The switching frequency of ICs is usually 1 MHz or higher. Therefore, as was shown, it is very important to study the microtransducer performance at the high operating frequency. This can be performed using Maxwell's equations, which will lead to the high-fidelity mathematical models [2].

Single-Phase Reluctance Micromotors: Microfabrication, Modeling, and Analysis

Consider the single-phase reluctance micromachined motors as illustrated in Fig. 20.129.

The emphases are concentrated on the analysis, modeling, and control of reluctance micromotors in the rotational microtransducer applications. Therefore, mathematical models must be found. The lumped-parameter modeling paradigm is based upon the use of the circuitry (voltage and current) and mechanical (velocity and displacement) variables to derive the differential equations using Newton's and Kirchhoff's laws. In these differential equations, the micromotor parameters are used. In particular, for the studied micromotor, the parameters are the stator resistance r_s, the magnetizing inductances in the *quadrature* and *direct* axes L_{mq} and L_{md}, the average magnetizing inductance $\bar{L}_m$, the leakage inductance L_{ls}, the moment of inertia J, and the viscous friction coefficient B_m.

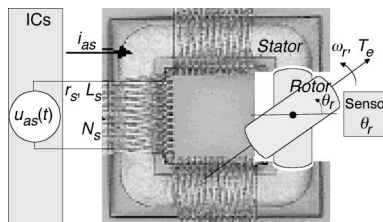



FIGURE 20.129 Single-phase reluctance micromotor with ICs and rotor displacement (position) sensor.

The expression for the electromagnetic torque was derived in [8]. In particular,

$$T_e = L_{\Delta m} i_{as}^2 \sin 2\theta_r$$

where $L_{\Delta m}$ is the half-magnitude of the sinusoidal magnetizing inductance L_m variations,

$$L_m(\theta_r) = \bar{L}_m - L_{\Delta m} \cos 2\theta_r$$

Thus, to develop the electromagnetic torque, the current i_{as} must be fed as a function of the rotor angular displacement θ_r. For example, if $i_{as} = i_M \mathrm{Re}(\sqrt{\sin 2\theta_r})$, then

$$T_{eaverage} = \frac{1}{\pi}\int_0^{\pi} L_{\Delta m} i_{as}^2 \sin 2\theta_r \, d\theta_r = \frac{1}{4} L_{\Delta m} i_M^2$$

The micromotor under our consideration is the synchronous micromachine, and the obtained expression for the phase current is very important to control the microtransducer. In particular, the Hall-effect position sensor should be used to measure the rotor displacement, and the ICs must feed the phase current as a nonlinear function of θ_r. Furthermore, the electromagnetic torque is controlled by changing the current magnitude i_M.

The mathematical model of the single-phase reluctance micromotor is found using Kirchhoff's and Newton's second laws. In particular, we have

$$u_{as} = r_s i_{as} + \frac{d\psi_{as}}{dt} \quad \text{(circuitry equation—Kirchhoff's law)}$$

$$T_e - B_m \omega_r - T_L = J\frac{d^2\theta_r}{dt^2} \quad \text{(\textit{torsional-mechanical} equation—Newton's law)}$$

Here, the electrical angular velocity ω_r and displacement θ_r are used as the mechanical system variables.

From $u_{as} = r_s i_{as} + \frac{d\psi_{as}}{dt}$ and the flux linkage equation $\psi_{as} = (L_{ls} + \bar{L}_m - L_{\Delta m} \cos 2\theta_r)_{as}$, using the *torsional-mechanical* dynamics, one obtains a set of three first-order nonlinear differential equations which models single-phase reluctance micromotors. In particular, we have

$$\frac{di_{as}}{dt} = -\frac{r_s}{L_{ls} + \bar{L}_m - L_{\Delta m}\cos 2\theta_r} i_{as} - \frac{2L_{\Delta m}}{L_{ls} + \bar{L}_m - L_{\Delta m}\cos 2\theta_r} i_{as}\omega_r \sin 2\theta_r + \frac{1}{L_{ls} + \bar{L}_m - L_{\Delta m}\cos 2\theta_r} u_{as}$$

$$\frac{d\omega_r}{dt} = \frac{1}{J}(L_{\Delta m} i_{as}^2 \sin 2\theta_r - B_m \omega_r - T_L)$$

$$\frac{d\theta_r}{dt} = \omega_r$$

As the mathematical model is found and the micromotor parameters are measured, nonlinear simulation and analysis can be straightforwardly performed to study the dynamic responses and analyze the micromotor efficiency. In particular, the resistance, inductances, moment of inertia, viscous friction coefficient, and other parameters can be directly measured or identified based upon micromotor testing. The steady-state and dynamic analysis based upon the lumped-parameter mathematical model is straightforward. However, the lumped-parameter mathematical models simplify the analysis, and thus, these models must be compared with the experimental data to validate the results.

The disadvantage of single-phase reluctance micromotors are high torque ripple, vibration, noise, low reliability, etc. Therefore, let us study three-phase synchronous reluctance micromotors.

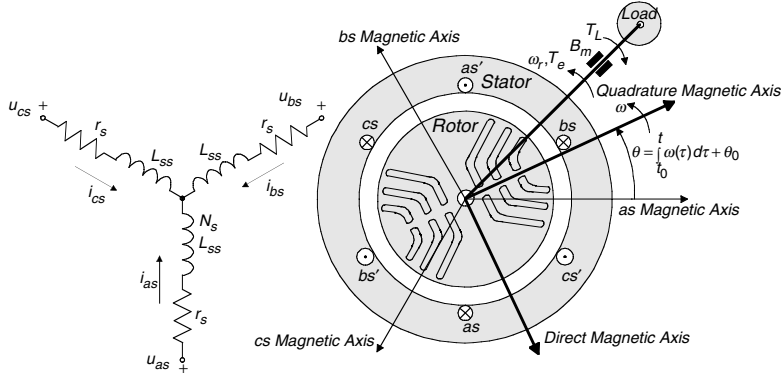



FIGURE 20.130 Three-phase synchronous reluctance micromotor.

Three-Phase Synchronous Reluctance Micromotors: Modeling and Analysis

Our goal is to address and solve a spectrum of problems in analysis, modeling, and control of synchronous reluctance micromachines. The electromagnetic features must be thoroughly analyzed before attempt to control micromotors. In fact, electromagnetic features significantly restrict the control algorithms to be applied. Depending upon the conceptual methods employed to analyze synchronous reluctance micromachines, different control laws can be designed and implemented using ICs. Analysis and control of synchronous reluctance micromotors can be performed using different modeling, analysis, and optimization concepts. Complete lumped-parameter mathematical models of synchronous reluctance micromotors in the *machine* (*abc*) and in the *quadrature, direct*, and *zero* (*qd*0) variables should be developed in the form of nonlinear differential equations. In particular, the circuitry lumped-parameters mathematical model is found using the Kirchhoff's voltage law. We have, see Fig. 20.130,

$$\mathbf{u}_{abcs} = \mathbf{r}_s \mathbf{i}_{abcs} + \frac{d\psi_{abcs}}{dt}$$

where u_{as}, u_{bs}, and u_{cs} are the phase voltages; i_{as}, i_{bs}, and i_{cs} are the phase currents; ψ_{as}, ψ_{bs}, and ψ_{cs} are the flux linkages,

$$\psi_{abcs} = \mathbf{L}_s \mathbf{i}_{abcs}, \qquad \mathbf{r}_s = \begin{bmatrix} r_s & 0 & 0 \\ 0 & r_s & 0 \\ 0 & 0 & r_s \end{bmatrix}$$

$$\mathbf{L}_s = \begin{bmatrix} L_{ls} + \bar{L}_m - L_{\Delta m}\cos(2\theta_r) & -\frac{1}{2}\bar{L}_m - L_{\Delta m}\cos 2\left(\theta_r - \frac{1}{3}\pi\right) & -\frac{1}{2}\bar{L}_m - L_{\Delta m}\cos 2\left(\theta_r + \frac{1}{3}\pi\right) \\ -\frac{1}{2}\bar{L}_m - L_{\Delta m}\cos 2\left(\theta_r - \frac{1}{3}\pi\right) & L_{ls} + \bar{L}_m - L_{\Delta m}\cos 2\left(\theta_r - \frac{2}{3}\pi\right) & -\frac{1}{2}\bar{L}_m - L_{\Delta m}\cos 2(\theta_r + \pi) \\ -\frac{1}{2}\bar{L}_m - L_{\Delta m}\cos 2\left(\theta_r + \frac{1}{3}\pi\right) & -\frac{1}{2}\bar{L}_m - L_{\Delta m}\cos 2(\theta_r + \pi) & L_{ls} + \bar{L}_m - L_{\Delta m}\cos 2\left(\theta_r + \frac{2}{3}\pi\right) \end{bmatrix}$$

$$\bar{L}_m = \frac{1}{3}(L_{mq} + L_{md}) \quad \text{and} \quad L_{\Delta m} = \frac{1}{3}(L_{md} - L_{mq})$$

The micromachine parameters are the stator resistance r_s, the magnetizing inductances in the *quadrature* and *direct* axes L_{mq} and L_{md}, the average magnetizing inductance $\bar{L}_m$, the leakage inductance L_{ls}, the moment of inertia J, and the viscous friction coefficient B_m.

The expressions for inductances are nonlinear functions of the electrical angular displacement θ_r. Hence, the *torsional-mechanical* dynamics must be used. Taking note of the Newton's second law of rotational motion, and using ω_r and θ_r (electrical angular velocity and displacement) as the state variables (mechanical variables), one obtains

$$T_e - B_m \frac{2}{P}\omega_r - T_L = J\frac{2}{P}\frac{d\omega_r}{dt}, \quad \frac{d\theta_r}{dt} = \omega_r$$

where T_e and T_L are the electromagnetic and load torques.

Torque Production Analysis—Using the coenergy, the electromagnetic torque, which is a nonlinear function of the micromotor variables (phase currents and electrical angular position) and micromotor parameters (number of poles P and inductance $L_{\Delta m}$), is found to be [8],

$$T_e = \frac{P}{2}L_{\Delta m}\left[i_{as}^2 \sin 2\theta_r + 2i_{as}i_{bs}\sin 2\left(\theta_r - \frac{1}{3}\pi\right) + 2i_{as}i_{cs}\sin 2\left(\theta_r + \frac{1}{3}\pi\right)\right.$$
$$\left. + i_{bs}^2 \sin 2\left(\theta_r - \frac{2}{3}\pi\right) + 2i_{bs}i_{cs}\sin 2\theta_r + i_{cs}^2 \sin 2\left(\theta_r + \frac{2}{3}\pi\right)\right]$$

To control the angular velocity, the electromagnetic torque must be regulated. To maximize the electromagnetic torque, ICs must feed the following phase currents as functions of the angular displacement measuring or observing (sensorless control) θ_r

$$i_{as} = \sqrt{2}i_M \sin\left[\theta_r + \frac{1}{3}\varphi_i\pi\right]$$

$$i_{bs} = \sqrt{2}i_M \sin\left[\theta_r - \frac{1}{3}(2-\varphi_i)\pi\right]$$

$$i_{cs} = \sqrt{2}i_M \sin\left[\theta_r + \frac{1}{3}(2+\varphi_i)\pi\right]$$

Thus, for $\varphi_i = 0.3245$, one obtains

$$T_e = \sqrt{2}PL_{\Delta m}i_M^2$$

That is, T_e is maximized and controlled by changing the magnitude of the phase currents i_M. Furthermore, it is no torque ripple (in practice, based upon the experimental results, and performing the high-fidelity modeling integrating nonlinear electromagnetics using Maxwell's equations, one finds that there exists the torque ripple which is due to the cogging torque, eccentricity, bearing, pulse-width-modulation, and other phenomena).

The majority of ICs are designed to control the phase voltages u_{as}, u_{bs}, and u_{cs}. Therefore, the three-phase balance voltage set is important. We have

$$u_{as} = \sqrt{2}u_M \sin\left[\theta_r + \frac{1}{3}\varphi_i\pi\right]$$

$$u_{bs} = \sqrt{2}u_M \sin\left[\theta_r - \frac{1}{3}(2-\varphi_i)\pi\right]$$

$$u_{cs} = \sqrt{2}u_M \sin\left[\theta_r + \frac{1}{3}(2+\varphi_i)\pi\right]$$

where u_M is the magnitude of the supplied voltages.

The mathematical model of synchronous reluctance micromotors in the *abc* variables is found to be

$$\frac{di_{as}}{dt} = \frac{1}{L_D}\Bigg\{(r_s i_{as} - u_{as})(4L_{ls}^2 + 3\bar{L}_m^2 - 3L_{\Delta m}^2 + 8\bar{L}_m L_{ls} - 4L_{ls}L_{\Delta m}\cos 2\theta_r)$$

$$+ (r_s i_{bs} - u_{bs})\left[3\bar{L}_m^2 - 3L_{\Delta m}^2 + 2\bar{L}_m L_{ls} + 4L_{ls}L_{\Delta m}\cos 2\left(\theta_r - \frac{1}{3}\pi\right)\right]$$

$$+ (r_s i_{cs} - u_{cs})\left[3\bar{L}_m^2 - 3L_{\Delta m}^2 + 2\bar{L}_m L_{ls} + 4L_{ls}L_{\Delta m}\cos 2\left(\theta_r + \frac{1}{3}\pi\right)\right]$$

$$+ 6\sqrt{3}L_{\Delta m}^2 L_{ls}\omega_r(i_{cs} - i_{bs}) + (8L_{\Delta m}L_{ls}^2\omega_r + 12L_{\Delta m}\bar{L}_m L_{ls}\omega_r)$$

$$\times\left(\sin 2\theta_r i_{as} + \sin 2\left(\theta_r - \frac{1}{3}\pi\right)i_{bs} + \sin 2\left(\theta_r + \frac{1}{3}\pi\right)i_{cs}\right)\Bigg\}$$

$$\frac{di_{bs}}{dt} = \frac{1}{L_D}\Bigg\{(r_s i_{as} - u_{as})\left[3\bar{L}_m^2 - 3L_{\Delta m}^2 + 2\bar{L}_m L_{ls} + 4L_{ls}L_{\Delta m}\cos 2\left(\theta_r - \frac{1}{3}\pi\right)\right]$$

$$+ (r_s i_{bs} - u_{bs})\left[4L_{ls}^2 + 3\bar{L}_m^2 - 3L_{\Delta m}^2 + 8\bar{L}_m L_{ls} - 4L_{ls}L_{\Delta m}\cos 2\left(\theta_r + \frac{1}{3}\pi\right)\right]$$

$$+ (r_s i_{cs} - u_{cs})[3\bar{L}_m^2 - 3L_{\Delta m}^2 + 2\bar{L}_m L_{ls} + 4L_{ls}L_{\Delta m}\cos 2\theta_r]$$

$$+ 6\sqrt{3}L_{\Delta m}^2 L_{ls}\omega_r(i_{as} - i_{cs}) + (8L_{\Delta m}L_{ls}^2\omega_r + 12L_{\Delta m}\bar{L}_m L_{ls}\omega_r)$$

$$\times\left(\sin 2\left(\theta_r - \frac{1}{3}\pi\right)i_{as} + \sin 2\left(\theta_r + \frac{1}{3}\pi\right)i_{bs} + \sin 2\theta_r i_{cs}\right)\Bigg\}$$

$$\frac{di_{cs}}{dt} = \frac{1}{L_D}\Bigg\{(r_s i_{as} - u_{as})\left[3\bar{L}_m^2 - 3L_{\Delta m}^2 + 2\bar{L}_m L_{ls} + 4L_{ls}L_{\Delta m}\cos 2\left(\theta_r + \frac{1}{3}\pi\right)\right]$$

$$+ (r_s i_{bs} - u_{bs})(3\bar{L}_m^2 - 3L_{\Delta m}^2 + 2\bar{L}_m L_{ls} + 4L_{ls}L_{\Delta m}\cos 2\theta_r)$$

$$+ (r_s i_{cs} - u_{cs})\left[4L_{ls}^2 + 3\bar{L}_m^2 - 3L_{\Delta m}^2 + 8\bar{L}_m L_{ls} - 4L_{ls}L_{\Delta m}\cos 2\left(\theta_r - \frac{1}{3}\pi\right)\right]$$

$$+ 6\sqrt{3}L_{\Delta m}^2 L_{ls}\omega_r(i_{bs} - i_{as}) + (8L_{\Delta m}L_{ls}^2\omega_r + 12L_{\Delta m}\bar{L}_m L_{ls}\omega_r)$$

$$\times\left[\sin 2\left(\theta_r + \frac{1}{3}\pi\right)i_{as} + \sin 2\theta_r i_{bs} + \sin 2\left(\theta_r - \frac{1}{3}\pi\right)i_{cs}\right]\Bigg\}$$

$$\frac{d\omega_r}{dt} = \frac{P^2}{4J}L_{\Delta m}\Bigg[i_{as}^2\sin 2\theta_r + 2i_{as}i_{bs}\sin 2\left(\theta_r - \frac{1}{3}\pi\right) + 2i_{as}i_{cs}\sin 2\left(\theta_r + \frac{1}{3}\pi\right)$$

$$+ i_{bs}^2\sin 2\left(\theta_r - \frac{2}{3}\pi\right) + 2i_{bs}i_{cs}\sin 2\theta_r + i_{cs}^2\sin 2\left(\theta_r + \frac{2}{3}\pi\right)\Bigg] - \frac{B_m}{J}\omega_r - \frac{P}{2J}T_L$$

$$\frac{d\theta_r}{dt} = \omega_r$$

Here

$$\bar{L}_m = \frac{1}{3}(L_{mq} + L_{md}), \quad L_{\Delta m} = \frac{1}{3}(L_{md} - L_{mq}) \quad \text{and} \quad L_D = L_{ls}(9L_{\Delta m}^2 - 4L_{ls}^2 - 12\bar{L}_m L_{ls} - 9\bar{L}_m^2)$$

The mathematical model can be simplified. In particular, in the rotor reference frame, we apply the Park transformation [8]

$$\mathbf{u}_{qd0s}^r = \mathbf{K}_s^r \mathbf{u}_{abcs}, \qquad \mathbf{i}_{qd0s}^r = \mathbf{K}_s^r \mathbf{i}_{abcs}, \qquad \psi_{qd0s}^r = \mathbf{K}_s^r \psi_{abcs}$$

$$\mathbf{K}_s^r = \frac{2}{3}\begin{bmatrix} \cos\theta_r & \cos\left(\theta_r - \frac{2}{3}\pi\right) & \cos\left(\theta_r + \frac{2}{3}\pi\right) \\ \sin\theta_r & \sin\left(\theta_r - \frac{2}{3}\pi\right) & \sin\left(\theta_r + \frac{2}{3}\pi\right) \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

where u_{qs}, u_{ds}, u_{0s}, i_{qs}, i_{ds}, i_{0s}, and ψ_{qs}, ψ_{ds}, ψ_{0s} are the *qd*0 voltages, currents, and flux linkages.

Using the circuitry and *torsional-mechanical* dynamics, one finds the following nonlinear differential equations to model synchronous reluctance micromotors in the rotor reference frame:

$$\frac{di_{qs}^r}{dt} = -\frac{r_s}{L_{ls} + L_{mq}} i_{qs}^r - \frac{L_{ls} + L_{md}}{L_{ls} + L_{mq}} i_{ds}^r \omega_r + \frac{1}{L_{ls} + L_{mq}} u_{qs}^r$$

$$\frac{di_{ds}^r}{dt} = -\frac{r_s}{L_{ls} + L_{md}} i_{ds}^r + \frac{L_{ls} + L_{mq}}{L_{ls} + L_{md}} i_{qs}^r \omega_r + \frac{1}{L_{ls} + L_{md}} u_{ds}^r$$

$$\frac{di_{0s}^r}{dt} = -\frac{r_s}{L_{ls}} i_{0s}^r + \frac{1}{L_{ls}} u_{0s}^r$$

$$\frac{d\omega_r}{dt} = \frac{3P^2}{8J}(L_{md} - L_{mq}) i_{qs}^r i_{ds}^r - \frac{B_m}{J}\omega_r - \frac{P}{2J}T_L$$

$$\frac{d\theta_r}{dt} = \omega_r$$

One can easily observe that this model is much simpler compared with the lumped-parameter mathematical model derived using the *abc* variables.

To attain the balanced operation, the *quadrature* and *direct* currents and voltages must be derived using the *direct* Park transformation $\mathbf{i}_{qd0s}^r = \mathbf{K}_s^r \mathbf{i}_{abcs}$, $\mathbf{u}_{qd0s}^r = \mathbf{K}_s^r \mathbf{u}_{abcs}$. Hence, the *qd*0 voltages u_{qs}^r, u_{ds}^r, and u_{0s}^r are found using the three-phase balance voltage set. In particular, we have

$$u_{qs}^r = \sqrt{2}u_M, \quad u_{ds}^r = 0, \quad u_{0s}^r = 0$$

We derived the mathematical models of three-phase synchronous reluctance micromotors. Based upon the differential equations obtained, nonlinear analysis can be performed, and the phase currents and voltages needed to guarantee the balance operating conditions can be found. The results reported can be straightforwardly used in nonlinear simulation.

Microfabrication Aspects

The fabrication of electromechanical microstructures and microtransducers can be made through deposition of the conductors (coils and windings), magnetic core, insulating layers, as well as other microstructures (movable and stationary members and their components). The order of the processes,

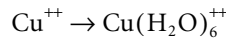
materials, and sequential steps are different depending on the MEMS which must be devised, designed, analyzed, and optimized first.

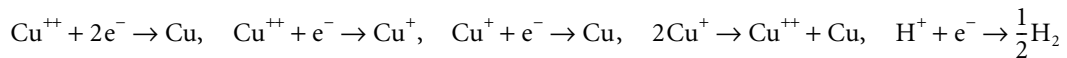
Conductor Thin Films Electrodeposition

The conductors (microcoils to make windings) in microstructures and microtransducers can be fabricated by electrodepositing the copper and other low resistivity metals. Electrodeposition of metals is made by immersing a conductive surface in a solution containing ions of the metal to be deposited. The surface is electrically connected to an external power supply, and current is fed through the surface into the solution. In general, the reaction of the metal ions ($Metal^{x+}$) with x electrons (xe^-) to form metal (Metal) is $Metal^{x+} + xe^- = Metal$.

To electrodeposit copper on the silicon wafer, the wafer is typically coated with a thin conductive layer of copper (seed layer) and immersed in a solution containing cupric ions. Electrical contact is made to the seed layer, and current is flowed (passed) such that the reaction $Cu^{2+} + 2e^- \rightarrow Cu$ occurs at the wafer surface. The wafer, which is electrically interacted such that the metal ions are changed to metal atoms, is the cathode. Another electrically active surface (anode) is the conductive solution to make the electrical path. At the anode, the oxidation reaction occurs that balances the current flow at the cathode, thus maintaining the electric neutrality. In the case of copper electroplating, all cupric ions removed from solution at the wafer cathode are replaced by dissolution from the copper anode. According to the Faraday law of electrolysis, in the absence of secondary reactions, the current delivered to a conductive surface during electroplating is proportional to the quantity of the metal deposited. Thus, the metal deposited can be controlled varying the electroplating current (current density) and the electrodeposition time.

The hydrated Cu ions reaction is

$$Cu^{++} \rightarrow Cu(H_2O)_6^{++}$$

and the cathode reactions are

$$Cu^{++} + 2e^- \rightarrow Cu, \quad Cu^{++} + e^- \rightarrow Cu^+, \quad Cu^+ + e^- \rightarrow Cu, \quad 2Cu^+ \rightarrow Cu^{++} + Cu, \quad H^+ + e^- \rightarrow \frac{1}{2}H_2$$

The copper electroplating solution commonly used is $CuSO_4–5H_2O$ (250 g/l) and H_2SO_4 (25 ml/l).

The basic processes are shown in Fig. 20.131, and the brief description of the sequential steps and equipment that can be used are given.

It must be emphasized that commonly used magnetic materials and conductors do not adhere well to silicon. Therefore, as was described, the adhesion layers (e.g., titanium Ti or chromium Cr) are deposited on the silicon surface prior to the magnetic material electroplating.

The electrodeposition rate is proportional to the current density and, therefore, the uniform current density at the substrate seed layer is needed to attain the uniform thickness of the electrodeposit. To achieve the selective electrodeposition, portions of the seed layer are covered with the resist (the current density at the mask edges nonuniform degrading electroplating). In addition to the current density, the deposition rate is also a nonlinear function of temperature, solution (chemicals), pH, direct/reverse current or voltage waveforms magnitude, waveform pulses, duty ratio, plating area, etc. In the simplest form, the thickness and electrodeposition time for the specified materials are calculated as

$$\text{Thickness}_{\text{material}} = \frac{\text{Time}_{\text{electroplating}} \times \text{Current}_{\text{density}} \times \text{Weight}_{\text{molecular}}}{\text{Faraday}_{\text{constant}} \times \text{Density}_{\text{material}} \times \text{Electron}_{\text{number}}}$$

$$\text{Time}_{\text{electroplating}} = \frac{\text{Thickness}_{\text{material}} \times \text{Faraday}_{\text{constant}} \times \text{Density}_{\text{material}} \times \text{Electron}_{\text{number}}}{\text{Current}_{\text{density}} \times \text{Weight}_{\text{molecular}}}$$

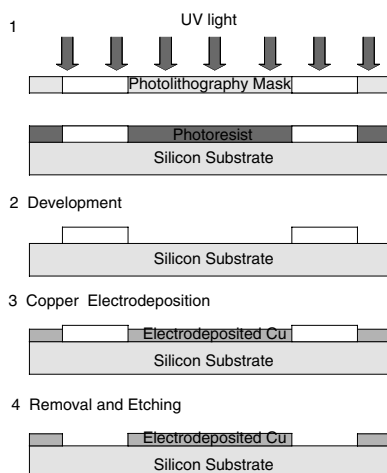



FIGURE 20.131 Electrodeposition of copper and basic processes: silicon, kapton, and other substrates, can be used. After clearing, the silicon substrate is covered with a 5–10 nm chromium or titanium and 100–200 nm copper seed layer by sputtering. The copper microcoils (microstructures) are patterned using the UV photolitography. The AZ-4562 photoresist can be spincoated and prebaked on a ramped hot plate at 90–100°C (ramp 30–40% with initial temperature 20–25°C) for one hour. Then, the photoresist is exposed in the Karl Suss Contact Masker with the energy 1200–1800 mJ cm^2. The development is released in 1:4 diluted alkaline solution (AZ-400) for 4–6 min. This gives the photoresist thickness 15–25 μm. Copper is electroplated with a three-electrode system with a copper anode and a saturated calomel reference electrode (the current power supply is the Perkin Elmer Current Source EG&G 263). The Shipley sulfate bath with the 5–10 ml/l brightener to smooth the deposit can be used. The electrodeposition is performed at 20–25°C with magnetic stirring and the dc current density 40–60 mA/cm^2 (this current density leads to smooth copper thin films with the 5–10 nm rms roughness for the 10 μm thickness of the deposited copper thin film). The resistivity of the electrodeposited copper thin film (microcoils) is 1.6–1.8 $\mu\Omega$ cm (close to the bulk copper resistivity). After the deposition, the photoresist is removed.

It was emphasized that electroplating is used to deposit thin-film conductors and magnetic materials. However, microtransducers need the insulation layers, otherwise the magnetic core and coils as well as multilayer microcoils themself will be short-circuited. Furthermore, the seed layers are embedded in microfabrication processes. As the magnetic core is fabricated on top of the microcoils (or microcoils are made on the magnetic core), the seed layer is difficult to remove because it is at the bottom or at the center of the microstructure. The mesh seed layer can serve as the electroplating seed layer for the lower conductors, and as the microstructure is made, the edges of the mesh seed layer can be exposed and removed through plasma etching [6]. Thus, the microcoils are insulated. It should be emphasized that relatively high aspect ratio techniques must be used to fabricate the magnetic core and microcoils, and patterning as well as surface planarization issues must be addressed.

NiFe Thin Films Electrodeposition

Magnetic cores in microstructures and microtransducers must be made. For example, the electroplated $Ni_{x\%}Fe_{100-x\%}$ thin films, such as *permalloy* $Ni_{80\%}Fe_{20\%}$, can be deposited to form the magnetic core of microtransducers (actuators and sensors), inductors, transformers, switches, etc. The basic processes and sequential steps used are similar to the processes for the copper electrodeposition and the electroplating is done in the electroplating bath. The windings (microcoils) must be insulated from magnetic cores, and therefore, the insulation layers must be deposited. The insulating materials used to insulate the windings from the magnetic core are benzocyclobutene, polyimide (PI-2611), etc. For example, the cyclotene 7200-35 is photosensitive and can be patterned through photolithography. The benzocyclobutene,

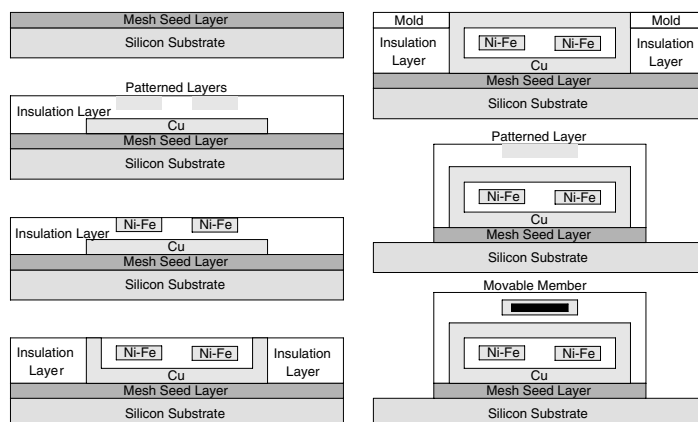



FIGURE 20.132 Basic fabrication sequential steps for the microtransducer fabrication.

used as the photoresist, offers good planarization and pattern properties, stability at low temperatures, and exhibits negligible hydrophilic properties.

The sketched fabrication process with sequential steps to make the electromagnetic microtransducer with a movable member is illustrated in Fig. 20.132. On the silicon substrate, the chromium–copper–chromium (Cr–Cu–Cr) mesh seed layer is deposited (through electron-beam evaporation) forming a seed layer for electroplating. The insulation layer (polyimide Dupont PI-2611) is spun on the top of the mesh seed layer to form the electroplating molds. Several coats can be done to obtain the desired thickness of the polyimide molds (one coat results in 8–12 μm insulation layer thickness). After coating, the polyimide is cured (at 280–310°C) in nitrogen for 1 h. A thin aluminum layer is deposited on top of the cured polyimide to form a hard mask for dry etching. Molds for the lower conductors are patterned and plasma etched until the seed layer is exposed. After etching the aluminum (hard mask) and chromium (top chromium–copper–chromium seed layer), the molds are filled with the electroplated copper, applying the described copper electroplating process. One coat of polyimide insulates the lower conductors and the magnetic core (thus, the insulation is achieved). The seed layer is deposited, mesh-patterned, coated with polyimide, and hard-cured. The aluminum thin layers (hard mask for dry etching) are deposited, and the mold for the magnetic cores is patterned and etched until the seed layer is exposed. After etching the aluminum (hard mask) and the chromium (top chromium–copper–chromium seed layer), the mold is filled with the electroplated $Ni_{x\%}Fe_{100-x\%}$ thin films (electroplating process). One coat of the insulating layer (polyimide) is spin-cast and cured to insulate the magnetic core and upper conductors. The via holes are patterned in the sputtered aluminum layer (hard mask) and etched through the polyimide layer using oxygen plasma. The vias are filled with the electroplated copper (electroplating process). A copper–chromium seed layer is deposited and the molds for the upper conductors are formed using thick photoresist. The molds are filled with the electroplated copper and removed. Then, the gap for the movable member is made using the conventional processes. After removing the seed layer, the passivation layer (polyimide) is coated and cured to protect the top conductors. The polyimide is masked and etched to the silicon substrate. The bottom mesh seed layer is wet etched and the microtransducer (with the ICs to control it) is diced and sealed.

Electroplated aluminum is the needed material to fabricate microstructures. In particular, aluminum can be used as the conductor to fabricate microcoils as well as mechanical microstructures (gears, bearing, pins, reflecting surfaces, etc.). Advanced techniques and processes for the electrodeposition of aluminum are documented in [9].

As was reported, the magnetic core of microstructures and microtransducers must be fabricated. Two major challenges in fabrication of high-performance microstructures are to make electroplated magnetic thin films with good magnetic properties as well as planarize microstructures (stationary and movable

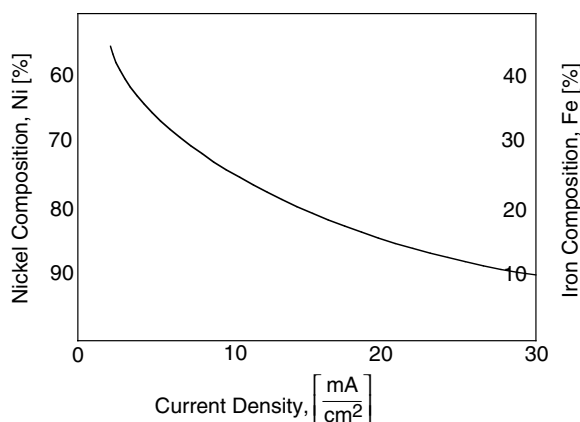



FIGURE 20.133 Nickel and iron compositions in $Ni_{x\%}Fe_{100-x\%}$ thin films as the functions of the current density.

members) [10]. Electroplating and micromolding techniques and processes are used to deposit NiFe alloys ($Ni_{x\%}Fe_{100-x\%}$ thin films), and $Ni_{80\%}Fe_{20\%}$ is called *permalloy*, while $Ni_{50\%}Fe_{50\%}$ is called *orthonol*.

Let us document the deposition process. To deposit $Ni_{x\%}Fe_{100-x\%}$ thin films, the silicon wafer is covered with a seed layer (for example, 15–25 nm chromium, 100–200 nm copper, and 25–50 nm chromium) deposited using electron beam evaporation. The photoresist layer (e.g., 10–20 μm Shipley STR-1110) is deposited on the seed layer and patterned. Then, the electrodeposition of the $Ni_{x\%}Fe_{100-x\%}$ is performed at 20–30°C using a two-electrode system, and the current density is in the range from 1 to 30 mA/cm^2. The temperature and pH should be maintained within the recommended values. High pH causes highly stressed NiFe thin films, and the low pH reduces leveling and cause chemical dissolving of the iron anodes resulting in disruption of the bath equilibrium and nonuniformity. High temperature leads to hazy deposits, and low temperature causes high current density burning. For deposition, the pulse-width-modulation (with varied waveforms, different forward and reverse magnitudes, and controlled duty cycle) can be used applying commercial or in-house made pulsed power supplies. Denoting the duty cycle length as T, the forward and reverse pulses lengths are denoted as T_f and T_r. The pulse length T can be 5–20 μs, and the duty cycle (ratio T_f/T_r) can be varied from 1 to 0.1. The ratio T_f/T_r influences the percentage of Ni in the $Ni_{x\%}Fe_{100-x\%}$ thin films, e.g., the composition of $Ni_{x\%}Fe_{100-x\%}$ can be regulated based upon the desired properties, which will be discussed later. However, varying the ratio T_f/T_r, the changes of the Ni are relatively modest (from 85% to 79%), and therefore, other parameters must vary to attain the desired composition.

It must be emphasized that the nickel (and iron) composition is a function of the current density, and Fig. 20.133 illustrates the nickel (iron) composition in the $Ni_{x\%}Fe_{100-x\%}$ thin films.

The $Ni_{80\%}Fe_{20\%}$ thin films of different thickness (which is a function of the electrodeposition time) are usually made at the current density 14–16 mA/cm^2. This range of the current density can be used to fabricate a various thickness of *permalloy* thin films (from 500 nm to 50 μm). The rms value of the thin film roughness is 4–7 nm for the 25 μm thickness. It should be emphasized that to guarantee good surface quality, the current density should be kept at the specified range, and usually to change the composition of the $Ni_{x\%}Fe_{100-x\%}$ thin films, the reverse current is controlled.

To attain a good deposit of the *permalloy*, the electroplating bath contains $NiSO_4$ (0.7 mol/l), $FeSO_4$ (0.03 mol/l), $NiCl_2$ (0.02 mol/l), saccharine (0.016 mol/l) as leveler (to reduce the residual stress allowing the fabrication of thicker films), and boric acid (0.4 mol/l).

The air agitation and saccharin were added to reduce internal stress and to keep the Fe composition stable. The deposition rate varies linearly as a function of the current density (the Faraday law is obeyed), and the electrodeposition slope is 100–150 nm cm^2/min mA. The *permalloy* thin films' density is 9 g/cm^3 (as for the bulk *permalloy*).

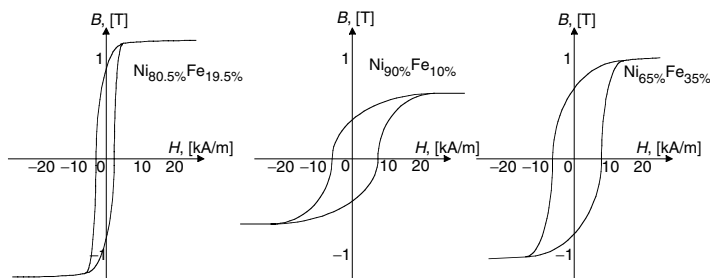



FIGURE 20.134 *B-H* curves for different $Ni_{x\%}Fe_{100-x\%}$ *permalloy* thin films.

The magnetic properties of the $Ni_{80\%}Fe_{20\%}$ (*permalloy*) thin films are studied, and the field coercivity (H_c) is a function of the thickness. For example, H_c = 650 A/m for 150 nm thickness and H_c = 30 A/m for 600 nm films.

Other $Ni_{80\%}Fe_{20\%}$ (deposited at 25°C) and $Ni_{50\%}Fe_{50\%}$ (deposited at 55°C) electroplating solutions are:

- $Ni_{80\%}Fe_{20\%}$: $NiSO_4$–$6H_2O$ (200 g/l/), $FeSO_4$–$7H_2O$ (9 g/l), $NiCl_2$–$6H_2O$ (5 g/l), H_3BO_3 (27 g/l), saccharine (3 g/l), and pH (2.5–3.5);
- $Ni_{50\%}Fe_{50\%}$: $NiSO_4$–$6H_2O$ (170 g/l/), $FeSO_4$–$7H_2O$ (80 g/l), $NiCl_2$–$6H_2O$ (138 g/l), H_3BO_3 (50 g/l), saccharine (3 g/l), and pH (3.5–4.5).

To electroplate $Ni_{x\%}Fe_{100-x\%}$ thin films, various additives and components (available from M&T Chemicals and other suppliers) can be used to control the internal stress and ductility of the deposit, keep the iron content solublized, obtain bright film and leveling of the process, attain the desired surface roughness, and most importantly to guarantee the desired magnetic properties.

In general, the *permalloy* thin films have optimal magnetic properties at the following composition: 80.5% of Ni and 19.5% of Fe. For $Ni_{80.5\%}Fe_{19.5\%}$ thin films, the material magnetostriction has zero crossing. Films with minimal magnetostriction usually have optimal coercivity and permeability properties, and, in general, the coercivity (depending on the films thickness) is 20 A/m (and higher as the thickness decreases), and permeability is from 600 to 2000. Varying the composition of Fe and Ni, the characteristics of the $Ni_{x\%}Fe_{100-x\%}$ thin films can be changed. The composition of the $Ni_{x\%}Fe_{100-x\%}$ thin films is controlled by changing the current density, T_f/T_r ratio (duty cycle), bath temperature (varying the temperature, the composition of Ni can be varied from 75 to 92%), reverse current (varying the reverse current in the range 0–1 A, the composition of Ni can be changed from 72 to 90%), air agitation of the solution, paddle frequency (0.1–1 Hz), forward and reverse pulses waveforms, etc. The *B-H* curves for three different $Ni_{x\%}Fe_{100-x\%}$ thin films are illustrated in Fig. 20.134. The $Ni_{80.5\%}Fe_{19.5\%}$ thin films have the saturation flux density 1.2 T, remanence B_r = 0.26 T-A/m, and the relative permeability 600–2000.

It must be emphasized that other electroplated permanent magnets (NiFeMo, NiCo, CoNiMnP, and other) and micromachined polymer magnets exhibit good magnetic properties and can be used as the alternative solution to the $Ni_{x\%}Fe_{100-x\%}$ thin films widely used.

NiFeMo and NiCo Thin Films Electrodeposition

To attain the desired magnetic properties (flux density, coercivity, permeability, etc.) and thickness, different thin film alloys can be used based upon the microstructure's and microtransducer's design, applications, and operating envelopes (temperature, shocks, radiation, humidity, etc.). As was discussed, the $Ni_{x\%}Fe_{100-x\%}$ thin films can be effectively used, and the desired magnetic properties can be readily achieved varying the composition of Ni and Fe. For sensors, the designer usually maximizes the flux density and permeability and minimizes the coercivity. The $Ni_{x\%}Fe_{100-x\%}$ thin films have the flux density up to 1.2 T, coercivity 20 (*permalloy*) to 500 A/m, and permeability 600–2000 (it was emphasized that

the magnetic properties also depend upon the thickness). Having emphasized the magnetic properties of the $Ni_{x\%}Fe_{100-x\%}$ thin films, let us perform the comparison. It was reported in the literature that [11]:

$Ni_{79\%}Fe_{17\%}Mo_{4\%}$ thin films have the flux density 0.7 T, coercivity 5 A/m, and permeability 3400, $Ni_{85\%}Fe_{14\%}Mo_{1\%}$ thin films have the flux density 1–1.1 T, coercivity 8–300 A/m, and permeability 3000–20000, $Ni_{50\%}Co_{50\%}$ thin films have the flux density 0.95–1.1 T, coercivity 1200–1500 A/m, and permeability 100–150 ($Ni_{79\%}Co_{21\%}$ thin films have the permeability 20).

In general, high flux density, low coercivity, and high permeability lead to high-performance MEMS. However, other issues (affordability, compliance, integrity, operating envelope, fabrication, etc.) must be also addressed while making the final choice. It must be emphasized that the magnetic characteristics, in addition to the film thickness, are significantly influenced by the fabrication processes and chemicals (materials) used.

The magnetic core in microstructures and microtransducers must be made. Two major challenges in fabrication of high-performance microstructures and microtransducers are to make electroplated magnetic thin films with good magnetic properties as well as planarize the stationary and movable members. Electroplating and micromolding techniques and processes are used to deposit NiFe, NiFeMo, and NiCo thin films. In particular, the $Ni_{80\%}Fe_{20\%}$, NiFeMo, and NiCo (deposited at 25°C) electroplating solutions are:

- $Ni_{80\%}Fe_{20\%}$: $NiSO_4$–$6H_2O$ (200 g/l), $FeSO_4$–$7H_2O$ (9 g/l), $NiCl_2$–$6H_2O$ (5 g/l), H_3BO_3 (27 g/l), and saccharine (3 g/l). The current density is 10–25 mA/cm^2 (nickel foil is used as the anode);
- NiFeMo: $NiSO_4$–$6H_2O$ (60 g/l/), $FeSO_4$–$7H_2O$ (4 g/l), Na_2MoO_4–$2H_2O$ (2 g/l), NaCl (10 g/l), citric acid (66 g/l), and saccharine (3 g/l). The current density is 10–30 mA/cm^2 (nickel foil is used as the anode);
- $Ni_{50\%}Co_{50\%}$: $NiSO_4$–$6H_2O$ (300 g/l/), $NiCl_2$–$6H_2O$ (50 g/l), $CoSO_4$–$7H_2O$ (30 g/l), H_3BO_3 (30 g/l), sodium lauryl sulfate (0.1 g/l), and saccharine (1.5 g/l). The current density is 10–25 mA/cm^2 (nickel or cobal can be used as the anode).

The most important feature is that the $Ni_{x\%}Fe_{100-x\%}$–NiFeMo–NiCo thin films (multiplayer nanocomposites) can be fabricated shaping the magnetic properties of the resulting materials to attain the desired performance characteristics through design and fabrication processes.

Micromachined Polymer Permanent Magnets

Electromagnetic microactuators can be deviced and fabricated using micromachined permanent magnet thin films including polymer magnets (magnetically hard ceramic ferrite powder imbedded in epoxy resin). Different forms and geometry of polymer magnets are available. Thin-film disks and plates are uniquely suitable for microactuator applications. For example, to actuate the mirrors in optical devices and optical MEMS, permanent magnets are used in rotational and translational (linear) microtransducers, microsensors, microswitches, etc. These polymer magnets have thickness ranging from hundreds of micrometers to several millimeters. Excellent magnetic properties can be achieved. For example, the micromachined polymer permanent-magnet disk with 80% strontium ferrite concentration (4 mm diameter and 90 μm thickness), magnetized normal to the thin-film plane (in the thickness direction), has the intrinsic coercivity H_{ci} = 320,000 A/m and a residual induction B_r = 0.06 T [12]. Permanent-magnet polymer magnets with thickness up to several millimeters can be fabricated by the low-temperature processes. To make the permanent magnets, the Hoosier Magnetics Co. strontium ferrite powder (1.1–1.5 μm grain size) and Shell epoxy resin (cured at 80°C for 2 h) can be used [12]. The polymer matrix contain a bisphenol-A-based epoxy resin diluted with cresylglycidyl ether and the aliphatic amidoamine is used as for curing. To prepare the polymer magnet composites, the strontium ferrite powder is mixed with the epoxy resin in the ball-mill rotating system (0.5 rad/s for many hours). After the aliphatic amidoamine is added, the epoxy is deposited and patterned using screen-printing. Then, the magnet is cured at 80°C for 2 h and magnetized in the desired direction.

It must be emphasized that magnets must be magnetized. That is, in addition to fabrication processes, one should study other issues, for example, the magnetization dynamics. The magnetic field in thin films are modeled, analyzed, and simulated solving differential equations, and the analytic and numerical results will be covered.

Magnetization Dynamics of Thin Films

The magnetic field, including the magnetization distribution, in thin films are modeled, analyzed, and simulated solving differential equations. The dynamic variables are the magnetic field density and intensity, magnetization, magnetization direction, wall position domain, etc. The thin films must be magnetized. Therefore, let us study the magnetization dynamics for thin films. To attain high-fidelity modeling, the magnetization dynamics in the angular coordinates is described by the Landay–Lifschitz–Gilbert equations [13]:

$$\frac{d\psi}{dt} = -\frac{\gamma}{M_s(1+\alpha^2)}\left(\sin^{-1}\psi\frac{\partial E(\theta,\psi)}{\partial\theta} + \alpha\frac{\partial E(\theta,\psi)}{\partial\psi}\right)$$

$$\frac{d\theta}{dt} = -\frac{\gamma\sin^{-1}\psi}{M_s(1+\alpha^2)}\left(\alpha\sin^{-1}\psi\frac{\partial E(\theta,\psi)}{\partial\theta} - \frac{\partial E(\theta,\psi)}{\partial\psi}\right)$$

where M_s is the saturation magnetization; $E(\theta, \psi)$ is the total Gibb's thin film free energy density; γ and α are the gyromagnetic and phenomenological constants.

The total energy consists the magnetocrystalline anisotropy energy, the exchange energy, and the magnetostatic self-energy (stray field energy) [14]:

$$E = \int_v\left(\frac{k_{\text{exh}}}{J_s^2}\sum_{j=1}^{3}(\nabla J_j)^2 - \frac{k_J}{J_s^2}(\mathbf{a}_J\,\mathbf{J})^2 - \frac{1}{2}\mathbf{J}\cdot\mathbf{H}_D - \mathbf{J}\cdot\mathbf{H}_{\text{ex}}\right)dv \quad \text{(Zeeman energy)}$$

$$\frac{\partial\mathbf{J}}{\partial t} = -|\gamma|\mathbf{J}\times\mathbf{H}_{\text{eff}} + \frac{\alpha}{J_s}\mathbf{J}\times\frac{\partial\mathbf{J}}{\partial t} \quad \text{(Gilbert equation)}$$

where $\mathbf{J}$ is the magnetic polarization vector, $\mathbf{H}_D$ and $\mathbf{H}_{\text{ex}}$ are the demagnetizing and external magnetic fields, $\mathbf{H}_{\text{eff}}$ is the effective magnetic field (sum of the applied, demagnetization, and anisotropy fields), k_{exh} and k_J are the exchange and magnetocrystalline anisotropy constants, $\mathbf{a}_J$ is the unit vector parallel to the uniaxial easy axis.

Using the vector notations, we have

$$\frac{d\mathbf{M}}{dt} = -\frac{\gamma}{1+\alpha^2}\left(\mathbf{M}\times\mathbf{H}_{\text{eff}} + \frac{\alpha}{M_s}\mathbf{M}\times(\mathbf{M}\times\mathbf{H}_{\text{eff}})\right)$$

Thus, using the nonlinear differential equations given, the high-fidelity modeling and analysis of nanostructured nanocomposite permanent magnets can be performed using field and material quantities, parameters, constants, etc.

Microstructures and Microtransducers with Permanent Magnets: Micromirror Actuator

The electromagnetic microactuator (permanent magnet on the cantilever flexible beam and spiral planar windings controlled by ICs fabricated using CMOS-MEMS technology) is illustrated in Fig. 20.135.

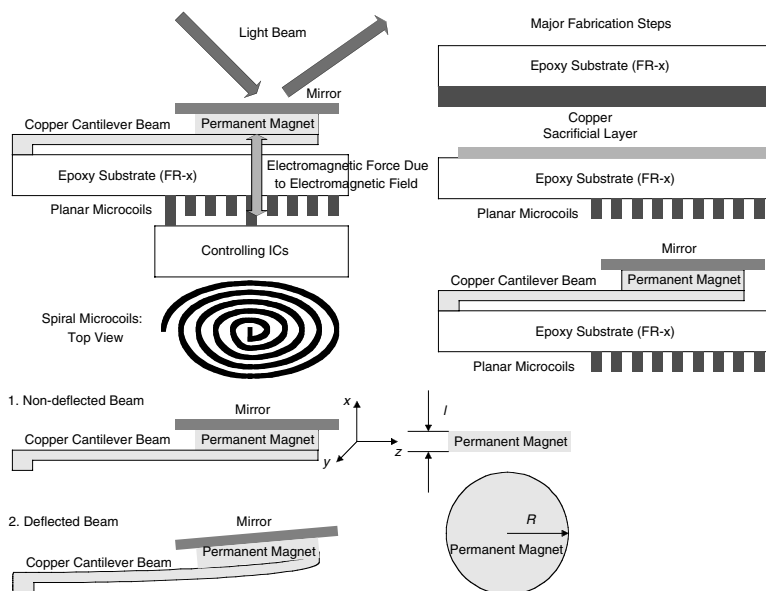



FIGURE 20.135 Electromagnetic microactuator with controlling ICs.

The electromagnetic microactuators can be made using conventional surface micromachining and CMOS fabrication technologies through electroplating, screen printing, lamination processes, sacrificial layer techniques, photolithography, etching, etc. In particular, the electromagnetic microactuator studied can be made on the commercially available epoxy substrates (e.g., FR series), which have the one-sided laminated copper layer (the copper layer thickness, which can be from 10 μm and higher, is defined by the admissible current density and the current value needed to establish the desired magnetic field to attain the specified mirror deflection, deflection rate, settling time, and other steady-state and dynamic characteristics). The spiral planar microcoils can be made on the one-sides laminated copper layer using photolithography and wet etching in the ferric chloride solution. The resulting x-μm thick N-turn microwinding will establish the magnetic field (the number of turns is a function of the footprint area available, thickness, spacing, outer-inner radii, geometry, fabrication techniques and processes used, etc.). After fabrication of the planar microcoils, the cantilever beam with the permanent magnet and mirror is fabricated on other side of the substrate. First, a photoresist sacrificial layer is spin-coated and patterned on the substrate. Then, a Ti–Cu–Cr seed layer is deposited to perform the copper electroplating (if the copper is used to fabricate the flexible cantilever structure). The second photoresist layer is spun and patterned to serve as a mold for the electroplating of the copper-based cantilever beam. The copper cantilever beam is electroplated in the copper-sulfate-based plating bath. After the electroplating, the photoresist plating mold and the seed layer are removed releasing the cantilever beam structure. It must be emphasized that depending upon the permanent magnet used, the corresponding fabricated processes must be done before or after releasing the beam. The permanent-magnet disk is positioned on the cantilever beam free end (for example, the polymer magnet can be screen-printed, and after curing the epoxy magnet, the magnet is magnetized by the external magnetic field). Then, the cantilever beam with the fabricated mirror is released by removing the sacrificial photoresist layer using acetone. It must be emphasized that the studied electromagnetic microactuator is fabricated using low-cost (affordable), high-yield micromachining—CMOS technology, processes, and materials. The most attractive feature is the application of the planar microcoils, which can be easily made. The use of the polymer permanent magnets (which have good magnetic properties) allows one to design high-performance electromagnetic microactuators. It must be emphasized that the polysilicon can be used to fabricate the cantilever beam and other permanent magnets can be applied.

In the article [12], the vertical electromagnetic force F_{ze}, acting on the permanent-magnet, is given by

$$F_{ze} = M_z \int_v \frac{dH_z}{dz} dv,$$

where M_z is the magenetization; H_z is the vertical component of the magnetic field intensity produced by the planar microwindings (H_z is a nonlinear function of the current fed or voltage applied to the microwindings, number of turns, microcoils, geometry, etc.; therefore, the thickness of the microcoils must be derived based on the maximum value of the current needed and the admissible current density).

The magnetically actuated cantilever microstructures were studied also in articles [15,16], and the expressions for the electromagnetic torque are found as the functions of the magnetic field using assumptions and simplifications which, in general, limit the applicability of the results. The differential equations which model the electromagnetic and *torsional-mechanical* dynamics can be derived. In particular, the equations for the electromagnetic field are found using electromagnetic theory, and the electromagnetic filed intensity H_z is controlled changing the current applied to the planar microwindings. The steady-state analysis, performed using the small-deflection theory [17], is also valuable. The static deflection of the cantilever beam x can be straightforwardly found using the force and beam quantities. In particular, $x = (l^3/3EJ)F_n$, where, l is the effective length of the beam; E is the Young's (elasticity) modulus; J is the equivalent moment of inertia of the beam with permanent magnet and mirror, and for the stand-alone cantilever beam with the rectangular cross section $J = \frac{1}{12}wh^3$; w and h are the width and thickness of the beam; F_n is the net force, which is normal to the cantilever beam.

In general, assuming that the magnetic flux is constant through the magnetic plane (loop), the torque on a planar current loop of any size and shape in the uniform magnetic field is

$$\mathbf{T} = i\mathbf{s} \times \mathbf{B} = \mathbf{m} \times \mathbf{B}$$

where i is the current and $\mathbf{m}$ is the magnetic dipole moment [A m^2].

Thus, the torque on the current loop always tends to turn the loop to align the magnetic field produced by the loop with the permanent-magnet magnetic field causing the resulting electromagnetic torque.

For example, for the current loop shown in Fig. 20.136, the torque (in N m) is found to be

$$\mathbf{T} = i\mathbf{s} \times \mathbf{B} = \mathbf{m} \times \mathbf{B} = 1 \times 10^{-3} \lfloor (1 \times 10^{-3})(2 \times 10^{-3})\mathbf{a}_z \rfloor \times (-0.5\mathbf{a}_y + \mathbf{a}_z) = 1 \times 10^{-9}\mathbf{a}_x$$

The electromagnetic force is found as

$$\mathbf{F} = \oint_l i d\mathbf{l} \times \mathbf{B}$$

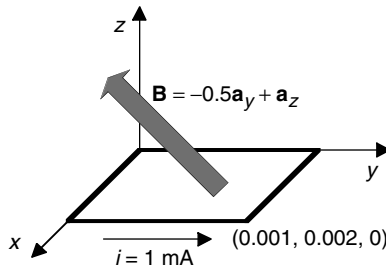



FIGURE 20.136 Rectangular planar loop in a uniform magnetic field with flux density $\mathbf{B} = -0.5\mathbf{a}_y + \mathbf{a}_z$.

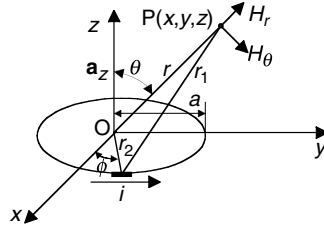



FIGURE 20.137 Planar current loop.

In general, the magnetic field quantities are derived using

$$\mathbf{B} = \frac{\mu_0}{4\pi} i \oint_l \frac{d\mathbf{l} \times r_0}{r^2} \quad \text{or} \quad \mathbf{H} = \frac{1}{4\pi} i \oint_l \frac{d\mathbf{l} \times r_0}{r^2}$$

and the Ampere circuital law gives

$$\oint_l \mathbf{H} \cdot d\mathbf{l} = i_{\text{total}} \quad \text{or} \quad \oint_l \mathbf{H} \cdot d\mathbf{l} = Ni$$

Making use of these expressions and taking note of the variables defined in Fig. 20.137, we have

$$\mathbf{H} = \frac{1}{4\pi} i \oint_l \frac{d\mathbf{l} \times r_1}{r_1^3} \quad \text{and} \quad \mathbf{B} = \frac{\mu_0}{4\pi} i \oint_l \frac{d\mathbf{l} \times \mathbf{r}_1}{r_1^3}$$

where $d\mathbf{I} = \mathbf{a}_\phi a\, d\phi = (-\mathbf{a}_x \sin\phi + \mathbf{a}_y \cos\phi)a\, d\phi$ and $\mathbf{r}_1 = \mathbf{a}_x\, (x - a\cos\phi) + \mathbf{a}_y\, (y - a\sin\phi) + \mathbf{a}_z\, z$.
Hence,

$$d\mathbf{I} \times \mathbf{r}_1 = [\mathbf{a}_x z\cos\phi + \mathbf{a}_y z\sin\phi - \mathbf{a}_z(y\sin\phi + x\cos\phi - a)]a\, d\phi.$$

Then, neglecting the small quantities ($a^2 \ll r^2$), we have

$$r_1^3 = (x^2 + y^2 + z^2 + a^2 - 2ax\cos\phi - 2ay\sin\phi)^{3/2} \approx r^3\left(1 - \frac{2ax}{r^2}\cos\phi - \frac{2ay}{r^2}\sin\phi\right)^{3/2}$$

Therefore, one obtains

$$\frac{1}{r_1^3} = \frac{1}{r^3}\left(1 + \frac{3ax}{r^2}\cos\phi + \frac{3ay}{r^2}\sin\phi\right)$$

Thus,

$$\mathbf{B} = \frac{\mu_0 a}{4\pi} i \int_0^{2\pi} [\mathbf{a}_x z\cos\phi + \mathbf{a}_y z\sin\phi - \mathbf{a}_z(y\sin\phi + x\cos\phi - a)]a\frac{1}{r^3}\left(1 + \frac{3ax}{r^2}\cos\phi + \frac{3ay}{r^2}\sin\phi\right)d\phi$$

$$= \frac{\mu_0 a^2}{4\pi r^3} i\left[\mathbf{a}_x \frac{3xz}{r^2} + \mathbf{a}_y \frac{3yz}{r^2} - \mathbf{a}_z\left(\frac{3x^2}{r^2} + \frac{3y^2}{r^2} - 2\right)\right]$$

Furthermore, using the coordinate transformation equations, in the spherical coordinate system one has

$$\mathbf{B} = \frac{\mu_0 a^2}{4\pi r^3} i(2\mathbf{a}_r \cos\theta + \mathbf{a}_\theta \sin\theta)$$

We have the expressions for the far-field components

$$B_r = \frac{\mu_0 a^2 \cos\theta}{2\pi r^3} i, \quad B_\theta = \frac{\mu_0 a^2 \sin\theta}{4\pi r^3} i, \quad B_\phi = 0$$

(due to the symmetry about the z axis, the magnetic flux density does not have the B_ϕ component).

Using the documented technique, one can easily find the magnetic vector potential. In particular, in general

$$\mathbf{A} = \frac{\mu_0}{4\pi} i \oint_l \frac{d\mathbf{l}}{r_1}$$

Assuming that $a^2 \ll r^2$, gives the following expression:

$$\frac{1}{r_1} = \frac{1}{r}\left(1 + \frac{ax}{r^2}\cos\phi + \frac{ay}{r^2}\sin\phi\right)$$

Therefore,

$$\mathbf{A} = \frac{\mu_0 a}{4\pi} i \int_0^{2\pi} (-\mathbf{a}_x \sin\phi + \mathbf{a}_y \cos\phi) \frac{1}{r}\left(1 + \frac{ax}{r^2}\cos\phi + \frac{ay}{r^2}\sin\phi\right) d\phi$$

$$= \frac{\mu_0 a}{4\pi r^3} i(-\mathbf{a}_x y + \mathbf{a}_y x)$$

Hence, in the spherical coordinate system, we obtain

$$\mathbf{A} = (\mathbf{A} \cdot \mathbf{a}_r)\mathbf{a}_r + (\mathbf{A} \cdot \mathbf{a}_\phi)\mathbf{a}_\phi + (\mathbf{A} \cdot \mathbf{a}_\theta)\mathbf{a}_\theta$$

$$= \frac{\mu_0 a}{4\pi r^2} i\mathbf{a}_\phi \sin\theta = A_\phi \mathbf{a}_\phi$$

It should be emphasized that the equations derived can be expressed using the magnetic dipole.

However, in the microtransducer studied, high-fidelity analysis should be performed. Hence, let us perform the comprehensive analysis.

The vector potential is found to be

$$A_\phi(r,\theta) = \frac{\mu_0 a i}{4\pi} \int_0^{2\pi} \frac{\cos\phi \, d\phi}{\sqrt{a^2 + r^2 - 2ar\sin\theta\cos\phi}}$$

and

$$B_r = \frac{1}{r\sin\theta} \frac{\partial(\sin\theta A_\phi)}{\partial\theta}, \quad B_\theta = -\frac{1}{r}\frac{\partial(rA_\phi)}{\partial r}, \quad B_\phi = 0$$

Making use of the following approximation:

$$A_\phi(r,\theta) = \frac{\mu_0 ai}{4\pi}\int_0^{2\pi} \frac{\cos\phi\, d\phi}{\sqrt{a^2 + r^2 - 2ar\sin\theta\cos\phi}} \approx \frac{\mu_0 a^2 r\sin\theta i}{4(a^2+r^2)^{3/2}}\left(1 + \frac{15a^2r^2\sin^2\theta}{8(a^2+r^2)^2} + \cdots\right)$$

one finds

$$B_r(r,\theta) = \frac{\mu_0 a^2\cos\theta i}{2(a^2+r^2)^{3/2}}\left(1 + \frac{15a^2r^2\sin^2\theta}{4(a^2+r^2)^2} + \cdots\right)$$

$$B_\theta(r,\theta) = -\frac{\mu_0 a^2\sin\theta i}{4(a^2+r^2)^{5/2}}\left(2a^2 - r^2 + \frac{15a^2r^2\sin^2\theta(4a^2-3r^2)}{8(a^2+r^2)^2} + \cdots\right)$$

$$B_\phi = 0$$

One can specify three regions:

- near the axis $\theta \ll 1$,
- at the center $r \ll a$,
- in far-field $r \gg a$.

The electromagnetic torque and field depend upon the current in the microwindings and are nonlinear functions of the displacement.

The expression for the electromagnetic forces and torques must be derived to model and analyze the *torsional-mechanical* dynamics. Newton's laws of motion can be applied to study the mechanical dynamics in the Cartesian or other coordinate systems (e.g., previously for the translational motion in the x-axis, we used

$$\frac{dv}{dt} = \frac{1}{m}(F_e - F_L) \quad \text{and} \quad \frac{dx}{dt} = v$$

to model the translational *torsional-mechanical* dynamics of the electromagnetic microactuators using the electromagnetic force F_e and the load force F_L).

For the studied microactuator, the rotational motion can be studied, and the electromagnetic torque can be approximated as

$$T_e = 4R^2 t_{tf} M H_p \cos\theta$$

where R and t_{tf} are the radius and thickness of the permanent-magnet thin-film disk; M is the permanent-magnet thin film magnetization; H_p is the field produced by the planar windings; θ is the displacement angle.

Then, the microactuator rotational dynamics is given by

$$\frac{d\omega}{dt} = \frac{1}{J}(T_e - T_L) \quad \text{and} \quad \frac{d\theta}{dt} = \omega$$

where T_L is the load torque, which integrates the friction and disturbances torques.

It should be emphasized that more complex and comprehensive mathematical models can be developed and used integrating the nonlinear electromagnetic and six-degree-of-freedom rotational-translational motions (*torsional-mechanical* dynamics) of the cantilever beam. As an illustration we consider the high-fidelity modeling of the electromagnetic system.

Electromagnetic System Modeling in Microactuators with Permanent Magnets: High-Fidelity Modeling and Analysis

In this section we focus our efforts to derive the expanded equations for the electromagnetic torque and force on cylindrical permanent-magnet thin films, see Fig. 20.135. The permanent-magnet thin film is assumed to be uniformly magnetized and the equations are developed for two orientations of the magnetization vector (the orientation is parallel to the axis of symmetry, and the orientation is perpendicular to this axis). Electromagnetic fields and gradients produced by the planar windings should be found at a point in inertial space, which coincides with the origin of the permanent-magnet axis system in its initial alignment. Our ultimate goal is to control microactuators, and thus, high-fidelity mathematical models (which will result in viable analysis, control, and optimization) must be derived. To attain our objective, the complete equations for the electromagnetic torque and force on a cylindrical permanent-magnet thin films are found.

The following notations are used: A, R, and l are the area, radius, and length of the cylindrical permanent magnet; $\mathbf{B}$ is the magnetic flux density vector; $\mathbf{B}_e$ is the expanded magnetic flux density vector; $[\partial\, \mathbf{B}]$ is the matrix of field gradients [T/m]; $[\partial\, \mathbf{B}_e]$ is the matrix of expanded field gradients [T/m]; $\mathbf{F}$ and $\mathbf{T}$ are the total force and torque vectors on the permanent-magnet thin film; i is the current in the planar microwinding; $\mathbf{m}$ is the magnetic moment vector [A m^2]; $\mathbf{M}$ is the magnetization vector [A/m]; $\mathbf{r}$ is the position vector (x, y, z are the coordinates in the Cartesian system),

$$\mathbf{r} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

T_r is the inertial coordinate vector-transformation matrix; W and Π are the work and potential energy; θ is the Euler orientation for the 3-2-1 rotation sequence; ∇ is the gradient operator; subscript ij represents partial derivative of i component in j-direction; subscript $(ij)k$ represents partial derivative of ij partial derivative in k-direction; $^{-}$(bar over a variable) indicates that it is referenced to the microactuator coordinates.

Electromagnetic Torques and Forces: Preliminaries

The equations for the electromagnetic torque and force on a cylindrical permanent-magnet thin film are found by integrating the equations for torques and forces on an incremental volume of the permanent-magnet thin film with magnetic moment $\mathbf{M}dv$ over the volume. Figure 20.135 illustrates the microactuator with the cylindrical permanent-magnet thin film in the coordinate system, which consists of a set of orthogonal body-fixed axes that are initially aligned with a set of orthogonal x-, y-, z-axes fixed in the inertial space.

The equations for the electromagnetic torque and force on an infinitesimal current can be derived using the fundamental relationship for the force on a current-carrying-conductor element in a uniform magnetic field. In particular, for a planar current loop (planar microwinding) with constant current i in the uniform magnetic field $\mathbf{B}$ (vector $\mathbf{B}$ gives the magnitude and direction of the flux density of the external field), the force on an element $d\mathbf{l}$ of the conductor is found using the Lorentz force law

$$\mathbf{F} = \oint_l i\, d\mathbf{l} \times \mathbf{B}$$

Assuming that the magnetic flux is constant through the magnetic loop, the torque on a planar current loop of any size and shape in the uniform magnetic field is

$$\mathbf{T} = i \oint_l \mathbf{r} \times (d\mathbf{l} \times \mathbf{B}) = i \oint_l \left((\mathbf{r} \cdot \mathbf{B})\, d\mathbf{l} - \mathbf{B} \oint_l \mathbf{r} \cdot d\mathbf{l} \right)$$

Using Stokes's theorem, one has

$$\mathbf{T} = i\left(\oint_s d\mathbf{A} \times \nabla(\mathbf{r} \cdot \mathbf{B}) - \mathbf{B}\oint_s (\nabla \times \mathbf{r}) \cdot d\mathbf{A}\right) = i\oint_s d\mathbf{A} \times \mathbf{B}$$

or

$$\mathbf{T} = i\mathbf{A} \times \mathbf{B} = \mathbf{m} \times \mathbf{B}$$

The electromagnetic torque **T** acts on the infinitesimal current loop in a direction to align the magnetic moment **m** with the external field **B**, and if **m** and **B** are misaligned by the angle θ, we have

$$\mathbf{T} = \mathbf{mB}\sin\theta$$

The incremental potential energy and work are found as

$$dW = d\Pi = \mathbf{T}\,d\theta = m\mathbf{B}\sin\theta\,d\theta \qquad \text{and} \qquad W = \Pi = -m\mathbf{B}\cos\theta = -\mathbf{m} \cdot \mathbf{B}$$

Using the electromagnetic force, we have

$$dW = -d\Pi = \mathbf{F} \cdot d\mathbf{r} = -\nabla\Pi \cdot d\mathbf{r}$$

and

$$\mathbf{F} = -\nabla\Pi = \nabla(\mathbf{m} \cdot \mathbf{B}) = (\mathbf{m} \cdot \nabla)\mathbf{B}$$

Coordinate Systems and Electromagnetic Field

The transformation from the inertial coordinates to the permanent-magnet coordinates is

$$\bar{\mathbf{r}} = T_r\mathbf{r} = \begin{bmatrix} \cos\theta_y\cos\theta_z & \cos\theta_y\sin\theta_z & -\sin\theta_y \\ \sin\theta_x\sin\theta_y\cos\theta_z - \cos\theta_x\sin\theta_z & \sin\theta_x\sin\theta_y\sin\theta_z + \cos\theta_x\sin\theta_z & \sin\theta_x\cos\theta_y \\ \cos\theta_x\sin\theta_y\cos\theta_z + \sin\theta_x\sin\theta_z & \cos\theta_x\sin\theta_y\sin\theta_z - \sin\theta_x\cos\theta_z & \cos\theta_x\cos\theta_y \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix}$$

$$\mathbf{r} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \quad \bar{\mathbf{r}} = \begin{bmatrix} \bar{x} \\ \bar{y} \\ \bar{z} \end{bmatrix}$$

We use the transformation matrix

$$T_r = \begin{bmatrix} \cos\theta_y\cos\theta_z & \cos\theta_y\sin\theta_z & -\sin\theta_y \\ \sin\theta_x\sin\theta_y\cos\theta_z - \cos\theta_x\sin\theta_z & \sin\theta_x\sin\theta_y\sin\theta_z + \cos\theta_x\sin\theta_z & \sin\theta_x\cos\theta_y \\ \cos\theta_x\sin\theta_y\cos\theta_z + \sin\theta_x\sin\theta_z & \cos\theta_x\sin\theta_y\sin\theta_z - \sin\theta_x\cos\theta_z & \cos\theta_x\cos\theta_y \end{bmatrix}$$

If the deflections are small, we have

$$T_{rs} = \begin{bmatrix} 1 & \theta_z & -\theta_y \\ -\theta_z & 1 & \theta_x \\ \theta_y & -\theta_x & 1 \end{bmatrix}$$

It should be emphasized that we use the 3-2-1 orthogonal transformation matrix for the *z-y-x* Euler rotation sequence, and θ_x, θ_y, θ_z are the rotation Euler angle about the *x*, *y*, and *z* axes.

The field **B** and gradients of **B** produced by the microcoils fixed in the inertial frame and expressed assuming that the electromagnetic fields can be described by the second-order Taylor series. Expanding **B** about the origin of the *x*, *y*, *z* system as a Taylor series, we have [18]

$$\mathbf{B}_e = \mathbf{B} + (\mathbf{r} \cdot \nabla)\mathbf{B} + \frac{1}{2}(\mathbf{r} \cdot \nabla)^2\mathbf{B}$$

or

$$B_{ei} = B_i + \frac{\partial B_i}{\partial \mathbf{r}}\mathbf{r} + \frac{1}{2}\mathbf{r}^T\frac{\partial^2 B_i}{\partial \mathbf{r}^2}\mathbf{r}$$

where

$$\frac{\partial B_i}{\partial \mathbf{r}} = \begin{bmatrix} \frac{\partial B_i}{\partial x} & \frac{\partial B_i}{\partial y} & \frac{\partial B_i}{\partial z} \end{bmatrix} \quad \text{and} \quad \frac{\partial^2 B_i}{\partial \mathbf{r}^2} = \begin{bmatrix} \frac{\partial \frac{\partial B_i}{\partial x}}{\partial x} & \frac{\partial \frac{\partial B_i}{\partial x}}{\partial y} & \frac{\partial \frac{\partial B_i}{\partial x}}{\partial z} \\ \frac{\partial \frac{\partial B_i}{\partial y}}{\partial x} & \frac{\partial \frac{\partial B_i}{\partial y}}{\partial y} & \frac{\partial \frac{\partial B_i}{\partial y}}{\partial z} \\ \frac{\partial \frac{\partial B_i}{\partial z}}{\partial x} & \frac{\partial \frac{\partial B_i}{\partial z}}{\partial y} & \frac{\partial \frac{\partial B_i}{\partial z}}{\partial z} \end{bmatrix}$$

We denote

$$B_{ij} = \frac{\partial B_i}{\partial j} \quad \text{and} \quad B_{(ij)k} = \frac{\partial \frac{\partial B_i}{\partial j}}{\partial k}$$

Then,

$$\frac{\partial B_i}{\partial \mathbf{r}} = [B_{ix} \quad B_{iy} \quad B_{iz}] \quad \text{and} \quad \frac{\partial^2 B_i}{\partial \mathbf{r}^2} = \begin{bmatrix} B_{(ix)x} & B_{(ix)y} & B_{(ix)z} \\ B_{(iy)x} & B_{(iy)y} & B_{(iy)z} \\ B_{(iz)x} & B_{(iz)y} & B_{(iz)z} \end{bmatrix}$$

Hence, the first-order gradients are given as

$$B_{eij} = B_{ij} + \frac{\partial \frac{\partial B_i}{\partial j}}{\partial \mathbf{r}}\mathbf{r} = B_{ij} + [B_{(ij)x} \quad B_{(ij)y} \quad B_{(ij)z}]\mathbf{r}$$

The expanded field is expressed in the permanent-magnet coordinates as

$$\bar{\mathbf{B}}_e = \bar{\mathbf{B}} + (\bar{\mathbf{r}} \cdot \bar{\nabla})\bar{\mathbf{B}} + \frac{1}{2}(\bar{\mathbf{r}} \cdot \bar{\nabla})^2\bar{\mathbf{B}}$$

where $\bar{\mathbf{B}} = T_r\mathbf{B}$ and $\bar{\nabla} = T_r\nabla$.

Using $\mathbf{r} = T_r^T \bar{\mathbf{r}}$, one has

$$B_{ei} = B_i + \frac{\partial B_i}{\partial \mathbf{r}} T_r^T \bar{\mathbf{r}} + \frac{1}{2} \bar{\mathbf{r}}^T T_r \frac{\partial^2 B_i}{\partial \mathbf{r}^2} T_r^T \bar{\mathbf{r}}$$

and

$$\bar{\mathbf{B}}_e = T_r \begin{bmatrix} B_x + \frac{\partial B_x}{\partial \mathbf{r}} T_r^T \bar{\mathbf{r}} + \frac{1}{2} \bar{\mathbf{r}}^T T_r \frac{\partial^2 B_x}{\partial \mathbf{r}^2} T_r^T \bar{\mathbf{r}} \\ B_y + \frac{\partial B_y}{\partial \mathbf{r}} T_r^T \bar{\mathbf{r}} + \frac{1}{2} \bar{\mathbf{r}}^T T_r \frac{\partial^2 B_y}{\partial \mathbf{r}^2} T_r^T \bar{\mathbf{r}} \\ B_z + \frac{\partial B_z}{\partial \mathbf{r}} T_r^T \bar{\mathbf{r}} + \frac{1}{2} \bar{\mathbf{r}}^T T_r \frac{\partial^2 B_z}{\partial \mathbf{r}^2} T_r^T \bar{\mathbf{r}} \end{bmatrix}$$

Electromagnetic Torques and Forces

Now let us derive the fields and gradients at any point in the permanent magnet using the second-order Taylor series approximation. To eliminate the transformations between the inertial and permanent magnet coordinate systems and simplify the second-order negligible small components, we assume that the relative motion between the magnet and the reference inertial coordinate is zero and the T_{rs} transformation matrix is used (otherwise, the second-order gradient terms will lead to cumbersome results).

The magnetization (the magnetic moment per unit volume) is constant over the volume of the permanent-magnet thin films, and $\mathbf{m} = \mathbf{M}v$.

Assuming that the magnetic flux is constant, the total electromagnetic torque and force on a planar current loop (microwinding) in the uniform magnetic field is

$$\bar{\mathbf{T}} = \int_v (\bar{\mathbf{M}} \times \mathbf{B}_e + \bar{\mathbf{r}} \times (\bar{\mathbf{M}} \cdot \nabla)\mathbf{B}_e) dv$$

$$\bar{\mathbf{F}} = \int_v (\bar{\mathbf{M}} \cdot \nabla)\mathbf{B}_e \, dv$$

where

$$(\bar{\mathbf{M}} \cdot \nabla)\mathbf{B}_e = [\partial \mathbf{B}_e]\bar{\mathbf{M}} = \begin{bmatrix} B_{exx} & B_{exy} & B_{exz} \\ B_{eyx} & B_{eyy} & B_{eyz} \\ B_{ezx} & B_{ezy} & B_{ezz} \end{bmatrix} \begin{bmatrix} M_{\bar{x}} \\ M_{\bar{y}} \\ M_{\bar{z}} \end{bmatrix}$$

Case 1: Magnetization Along the Axis of Symmetry

For orientation of the magnetization vector along the axis of symmetry (x-axis) of the permanent-magnet thin films, we have

$$(\bar{\mathbf{M}} \cdot \nabla)\mathbf{B}_e = [\partial \mathbf{B}_e]\bar{\mathbf{M}} = M_{\bar{x}} \begin{bmatrix} B_{exx} \\ B_{exy} \\ B_{exz} \end{bmatrix}$$

Thus, in the expression $\bar{\mathbf{T}} = \int_v (\bar{\mathbf{M}} \times \mathbf{B}_e + \bar{\mathbf{r}} \times (\bar{\mathbf{M}} \cdot \nabla)\mathbf{B}_e) dv$,

the terms are

$$\bar{\mathbf{r}} \times (\overline{\mathbf{M}} \cdot \nabla)\mathbf{B}_e = M_{\bar{x}} \begin{bmatrix} -B_{exy}\bar{z} + B_{exz}\bar{y} \\ B_{exx}\bar{z} - B_{exz}\bar{x} \\ -B_{exx}\bar{y} + B_{exy}\bar{x} \end{bmatrix} \quad \text{and} \quad \overline{\mathbf{M}} \times \mathbf{B}_e = M_{\bar{x}} \begin{bmatrix} 0 \\ -B_{ez} \\ B_{ey} \end{bmatrix}$$

Therefore,

$$T_{\bar{x}} = M_{\bar{x}} \int_v (B_{exz}\bar{y} - B_{exy}\bar{z})dv$$

$$T_{\bar{y}} = -M_{\bar{x}} \int_v B_{ez}\, dv + M_{\bar{x}} \int_v (B_{exx}\bar{z} - B_{exz}\bar{x})dv$$

and

$$T_{\bar{z}} = M_{\bar{x}} \int_v B_{ey}\, dv + M_{\bar{x}} \int_v (B_{exy}\bar{x} - B_{exx}\bar{y})dv$$

The terms in the derived equations must be evaluated.

Let us find the analytic expression for the electromagnetic torque $T_{\bar{x}}$. In particular, we have

$$\int_v B_{exz}\bar{y}\; dv = B_{xz}\int_v \bar{y}\; dv + B_{(xx)z}\int_v \bar{x}\bar{y}\; dv + B_{(xy)z}\int_v \bar{y}^2 dv + B_{(xz)z}\int_v \bar{z}\bar{y}\; dv$$

where

$$\int_v \bar{y}\; dv = 0, \quad \int_v \bar{x}\bar{y}\; dv = 0, \quad \int_v \bar{z}\bar{y}\; dv = 0$$

and

$$\int_v \bar{y}^2 dv = \int_{-\frac{1}{2}l}^{\frac{1}{2}l} \int_{-R}^{R} \int_{-\sqrt{R^2 - \bar{z}^2}}^{\sqrt{R^2 - \bar{z}^2}} \bar{y}^2\, d\bar{y}\, d\bar{z}\, d\bar{x} = \frac{1}{4}\pi l R^4 = \frac{1}{4} v R^4$$

Therefore,

$$M_{\bar{x}} \int_v B_{exz}\bar{y}\, dv = M_{\bar{x}}\frac{1}{4}B_{(xy)z} v R^4$$

Furthermore,

$$M_{\bar{x}} \int_v B_{exy}\bar{z}\, dv = M_{\bar{x}}\frac{1}{4}B_{(xy)z} v R^4$$

Thus, for $T_{\bar{x}}$, one has

$$T_{\bar{x}} = M_{\bar{x}} \int_v (B_{exz}\bar{y} - B_{exy}\bar{z})dv = M_{\bar{x}}\left(\frac{1}{4}B_{(xy)z} v R^4 - \frac{1}{4}B_{(xy)z} v R^4\right) = 0$$

Then, for $T_{\bar{y}}$, we obtain

$$\begin{aligned} T_{\bar{y}} &= -M_{\bar{x}}\int_v B_{ez}\,dv + M_{\bar{x}}\int_v (B_{exx}\bar{z} - B_{exz}\bar{x})\,dv \\ &= M_{\bar{x}}\left[-\left(B_z + B_{(zx)x}\frac{1}{24}l^2 + B_{(zy)y}\frac{1}{8}R^2 + B_{(zz)z}\frac{1}{8}R^2\right)v + B_{(xx)z}\left(\frac{1}{4}R^2 - \frac{1}{12}l^2\right)v\right] \\ &= -vM_{\bar{x}}\left[B_z + B_{(xx)z}\left(\frac{1}{4}R^2 - \frac{1}{8}l^2\right) + B_{(yy)z}\frac{1}{8}R^2 + B_{(zz)z}\frac{1}{4}R^2\right] \end{aligned}$$

Finally, we obtain the expression for T_z as

$$\begin{aligned} T_{\bar{z}} &= M_{\bar{x}}\int_v B_{ey}\,dv + M_{\bar{x}}\int_v (B_{exy}\bar{x} - B_{exx}\bar{y})\,dv \\ &= vM_{\bar{x}}\left[B_y + B_{(xx)y}\left(\frac{1}{8}l^2 - \frac{1}{4}R^2\right) - B_{(yy)y}\frac{1}{8}R^2 - B_{(yz)z}\frac{1}{8}R^2\right] \end{aligned}$$

Thus, the following electromagnetic torque equations result:

$$\begin{aligned} T_{\bar{x}} &= 0 \\ T_{\bar{y}} &= -vM_{\bar{x}}\left[B_z + B_{(xx)z}\left(\frac{1}{4}R^2 - \frac{1}{8}l^2\right) + B_{(yy)z}\frac{1}{8}R^2 + B_{(zz)z}\frac{1}{4}R^2\right] \\ T_{\bar{z}} &= vM_{\bar{x}}\left[B_y + B_{(xx)y}\left(\frac{1}{8}l^2 - \frac{1}{4}R^2\right) - B_{(yy)y}\frac{1}{8}R^2 - B_{(yz)z}\frac{1}{8}R^2\right] \end{aligned}$$

The electromagnetic forces are found as well. In particular, from

$$\begin{aligned} F_{\bar{x}} &= M_{\bar{x}}\int_v B_{exx}\,dv \\ F_{\bar{y}} &= M_{\bar{x}}\int_v B_{exy}\,dv \end{aligned}$$

and

$$F_{\bar{z}} = M_{\bar{x}}\int_v B_{exz}\,dv$$

using the expressions for the expanded magnetic fluxes, e.g.,

$$\int_v B_{exx}\,dv = \int_v (B_{xx} + B_{(xx)x}\bar{x} + B_{(xx)y}\bar{y} + B_{(xx)z}\bar{z})\,dv$$

and performing the integration, one has the following expressions for the electromagnetic forces as the function of the magnetic field:

$$F_{\bar{x}} = vM_{\bar{x}}B_{xx}, \quad F_{\bar{y}} = vM_{\bar{x}}B_{xy}, \quad F_{\bar{z}} = vM_{\bar{x}}B_{xz}$$

Case 2: Magnetization Perpendicular to the Axis of Symmetry

For orientation of the magnetization vector perpendicular to the axis of symmetry, the following equation is used to find the electromagnetic torque:

$$\bar{\mathbf{T}} = \int_v (\overline{\mathbf{M}} \times \mathbf{B}_e + \bar{\mathbf{r}} \times (\overline{\mathbf{M}} \cdot \nabla)\mathbf{B}_e)\, dv$$

where

$$(\overline{\mathbf{M}} \cdot \nabla)\mathbf{B}_e = [\partial \mathbf{B}_e]\overline{\mathbf{M}} = M_{\bar{z}} \begin{bmatrix} B_{exz} \\ B_{eyz} \\ B_{ezz} \end{bmatrix} \qquad \bar{\mathbf{r}} \times (\overline{\mathbf{M}} \cdot \nabla)\mathbf{B}_e = M_{\bar{z}} \begin{bmatrix} -B_{eyz}\bar{z} + B_{ezz}\bar{y} \\ B_{exz}\bar{z} - B_{ezz}\bar{x} \\ B_{exz}\bar{y} + B_{eyz}\bar{x} \end{bmatrix}$$

and

$$\overline{\mathbf{M}} \times \mathbf{B}_e = M_{\bar{z}} \begin{bmatrix} -B_{ey} \\ B_{ex} \\ 0 \end{bmatrix}$$

Thus,

$$\begin{aligned} T_{\bar{x}} &= -M_{\bar{z}} \int_v B_{ey}\, dv + M_{\bar{z}} \int_v (B_{exz}\bar{y} - B_{eyz}\bar{z})\, dv \\ T_{\bar{y}} &= M_{\bar{z}} \int_v B_{ex}\, dv + M_{\bar{z}} \int_v (B_{ezz}\bar{z} - B_{ezz}\bar{x})\, dv \\ T_{\bar{z}} &= M_{\bar{z}} \int_v (B_{eyz}\bar{x} - B_{exz}\bar{y})\, dv \end{aligned}$$

Expressing the fluxes and performing the integration, we have the following expressions for the torque components as the function of the magnetic field:

$$\begin{aligned} T_{\bar{x}} &= -vM_{\bar{z}}\left(B_y + B_{(xx)y}\frac{1}{24}l^2 + B_{(yy)y}\frac{1}{8}R^2 + B_{(yz)z}\frac{1}{8}R^2\right) \\ T_{\bar{y}} &= vM_{\bar{z}}\left[B_x + B_{(xz)y}\left(\frac{3}{8}R^2 - \frac{1}{12}l^2\right) + B_{(xx)x}\frac{1}{24}l^2 + B_{(xy)y}\frac{1}{8}R^2\right] \\ T_{\bar{z}} &= vM_{\bar{z}}B_{(xy)y}\left(\frac{1}{12}l^2 - \frac{1}{4}R^2\right) \end{aligned}$$

The electromagnetic forces are found to be

$$\begin{aligned} F_{\bar{x}} &= M_{\bar{z}} \int_v B_{exz}\, dv = vM_{\bar{z}}B_{xz} \\ F_{\bar{y}} &= M_{\bar{z}} \int_v B_{eyz}\, dv = vM_{\bar{z}}B_{yz} \\ F_{\bar{z}} &= M_{\bar{z}} \int_v B_{ezz}\, dv = vM_{\bar{z}}B_{zz} \end{aligned}$$

Thus, the expressions for the electromagnetic force and torque components are derived. These equations provide one with a clear perspective on how to model, analyze, and control the electromagnetic forces and torques changing the applied magnetic field because the terms

$$B_{ij} = \frac{\partial B_i}{\partial j} \quad \text{and} \quad B_{(ij)k} = \frac{\partial \frac{\partial B_i}{\partial j}}{\partial k}$$

can be viewed as the control variables. It must be emphasized that the electromagnetic field (B_{ij} and $B_{(ij)k}$) is controlled by regulating the current in the planar microwindings and designing the microwindings (or other radiating energy microdevices). As was discussed, the derived forces and torques must be used in the torsional-mechanical equations of motion for the microactuator, and, in general, the six-degree-of-freedom microactuator mechanical dynamics results. These mechanical equations of motion are easily integrated with the derived electromagnetic equations, and closed-loop systems can be designed to attain the desired microactuator performance. These equations guide us to the importance of electromagnetic features in the modeling, analysis, and design of microactuators.

Some Other Aspects of Microactuator Design and Optimization

In addition to the electromagnetic-mechanical (electromechanical) analysis and design, other design and optimization problems are involved. As an example, let us focus our attention on the planar windings. The ideal planar microwindings must produce the maximum electromagnetic field, minimizing the footprint area, taking into consideration the material characteristics, operating conditions, applications, power requirements, and many other factors. Many planar winding parameters and characteristics can be optimized, for example, the dc resistance must be minimized to improve the efficiency, increase the flux, decrease the losses, etc. To attain good performance, in general, microwindings have the concentric circular current path and no interconnect resistances. For N-turn winding, the total dc resistance r_t is found to be

$$r_t = \frac{2\pi\rho}{t_w} \sum_{k=1}^{N} \frac{1}{\ln(r_{Ok}/r_{Ik})}$$

where ρ is the winding material resistivity, t_w is the winding thickness, r_{Ok} and r_{Ik} are the outer and inner radii of the k-turn winding, respectively.

To achieve the lowest resistance, the planar winding radii can be optimized by minimizing the resistance, and the minimum resistance is denoted as $r_{t\min}$. In particular, making use of first- and second-order necessary conditions for minimization, one has

$$\frac{dr_t}{dr_w} = 0 \quad \text{and} \quad \frac{d^2 r_t}{dr_w^2} > 0$$

where r_w is the inner or outer radius of an arbitrary turn of the optimized planar windings from the standpoint of minimizing the resistance.

Then, the minimum value of the microcoil resistance is given by

$$r_{t\min} = \frac{2\pi\rho}{t_w} \frac{N}{(r_{OR}/r_{IR})}$$

where r_{OR} and r_{IR} are the outer and inner radii of the windings (i.e., $r_{O\ N\text{th microcoil}}$ and $r_{I\ \text{1st microcoil}}$), respectively.

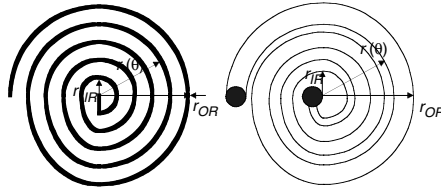



FIGURE 20.138 Planar spiral microwinding.

Thus, using the number of turns and turn-to-turn spacing, the outer and inner radii of the *k*-turn winding are found as

$$\frac{r_{Ok}}{r_{Ik}} = \left(\frac{r_{OR}}{r_{IR}}\right)^{1/N}$$

For spiral windings, the *averaging* (*equivalency*) concept should be used because the outer and inner radii are the functions of the planar angle, see Fig. 20.138. Finally, it should be emphasized that the width of the *N*th microcoil is specified by the rated voltage current density versus maximum current density needed, fabrication technologies used, material characteristics, etc.

Micromachined Polycrystalline Silicon Carbide Micromotors

Articles [19,20] report the silicon-based fabrication of reluctance micromotors. This section is focused on a new enabling technology to fabricate microtransducers. Multilayer fabrication processes at low temperature and micromolding techniques were developed to fabricate SiC microstructures and salient-pole micromotors, which can be used at a very high temperature (400°C and higher) [21,22]. This was done through the SiC surface micromachining. Advantages of the SiC micromachining and SiC technologies (high temperature and ruggedness) should be weighted against fabrication drawbacks because new processes must be designed and optimized. Reactive ion etching is used to pattern SiC thin films; however, many problems, such as masking, low etch rates, and poor etch selectivity, must be addressed and resolved. Articles [21,22] report two single-layer reactive ion etching-based polycrystalline SiC surface micromachining processes using polysilicon or SiO_2 as the sacrificial layer. In addition, the micromolding process, used to fabricated polysilicon molds in conjunction with polycrystalline SiC film deposition and mechanical polishing to pattern polycrystaline SiC films, are introduced. The micromolding process can be used for single- and multilayer SiC surface micromachining.

The micromotor fabrication processes are illustrated in Fig. 20.139. A 5–10 μm thick sacrificial molding polysilicon is deposited through the LPCVD on a 3–5 μm sacrificial thermal oxide. The rotor-stator mold formation can be made on the polished (chemical-mechanical polishing) polysilicon surface, enabling the 2 μm fabrication features using standard lithography and reactive ion etching. After the mold formation and delineation, the SiC is deposited on the wafer using atmospheric pressure chemical vapor deposition reactor. In particular, the phosphorus-doped (*n*-type) polycrystalline SiC films are deposited on the SiO_2 sacrificial layers at 1050°C with 0.5–1 μm/h rate (deposition is not selective, and SiC will be deposited on the surfaces of the polysilicon molds as well). Mechanical polishing of SiC is needed to expose the polysilicon and planarize the wafer surface (in [21,22], the polishing was done with 3 μm diameter diamond suspension, 360 N normal force, and 15 rad/sec pad rotation—the removal rate of SiC was reported to be 100 nm/min). The wafers are polished until the top surface of the polysilicon mold is exposed (polishing must be stopped at once due to the fast polishing rate). The flange mold is fabricated through the polysilicon and the sacrificial oxide etching (using the KOH and BHF, respectively). The 0.5 μm bearing clearance low-temperature oxide is deposited and annealed at 1000°C. Then, the 1 μm polycrystalline SiC film is deposited and patterned by reactive ion etching to make the bearing.

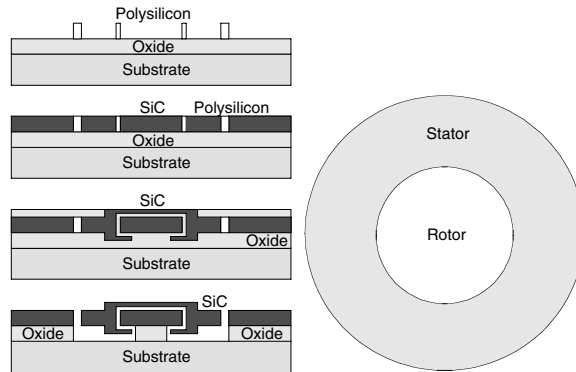



FIGURE 20.139 Fabrication of the SiC micromotors: cross-sectional schematics.

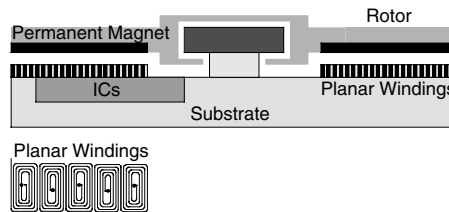



FIGURE 20.140 Slotless axial electromagnetic micromotor (cross-sectional schematics) with controlling ICs.

The release begins with the etching (BHF solution) to strip the left-over bearing clearance oxide. The sacrificial mold is removed by etching (KOH system) the polysilicon. It should be emphasized that the SiC and SiO are not etched during the mold removal step. Then, the moving parts of the micromotor were released. The micromotor is rinsed in water and methanol, and dried with the air jet.

Using this fabrication process, the micromotor with the 100–150 μm rotor diameter, 2 μm airgap, and 21 μm bearing radius, was fabricated and tested in [21, 22]. The rated voltage was 100 V and the maximum angular velocity was 30 rad/s. For silicon and polysilicon micromotors, two of the most critical problems are the bearing and ruggedness. The application of SiC reduces the friction and improves the ruggedness. These contribute to the reliability of the SiC-based fabricated micromachines.

Axial Electromagnetic Micromotors

The major problem is to devise novel microtransducers in order to eliminate fabrication difficulties and guarantee affordability, efficiency, reliability, and controllability of MEMS. In fact, the electrostatic and planar micromotor fabricated and tested to date are found to be inadequate for a wide range of applications due to difficulties associated and the cost. Therefore, this section is devoted to devising novel affordable rotational micromotors.

Figure 20.140 illustrates the devised axial topology micromotor, which has the *closed-ended* electromagnetic system. The stator is made on the substrate with deposited microwindings (printed copper coils can be made using the fabrication processes described as well as using a double-sided substrate with one-sided deposited copper thin films through conventional photolithography processes). The bearing post is fabricated on the stator substrate and the bearing hold is a part of the rotor microstructure. The rotor with permanent-magnet thin films rotates due to the electromagnetic torque developed. It is important to emphasize that the stator and rotor are made using conventional well-developed processes and materials.

It is evident that conventional silicon and SiC technologies can be used. The documented micromotor has a great number of advantages. The most critical benefit is the fabrication simplicity. In fact, axial

micromotors can be straightforwardly fabricated and this will enable their wide applications as microactuators and microsensors. However, the axial micromotors must be designed and optimized to attain good performance. The optimization is based upon electromagnetic, mechanical, and thermal design. The micromotor optimization can be carried out using the steady-state concept (finite element analysis) and dynamic paradigms (lumped-parameters models or complete electromagnetic-mechanical-thermal high-fidelity mathematical models derived as a set of partial differential equations using Maxwell's, *torsional-mechanical*, and heat equations). In general, the nonlinear optimization problems are needed to be addressed, formulated, and solved to guarantee the superior microtransducer performance. In addition to the microtransducer design, one must concentrate the attention on the ICs and controller design. In particular, the circuitry is designed based upon the converter and inverter topologies (e.g., hard- and soft-switching, one-, two-, or four-quadrant, etc.), filters and sensors used, rated voltage and current, etc. From the control prespective, the electromagnetic features must be thoroughly examined. For example, the electromagnetic micromotor studied is the synchronous micromachine. Therefore, to develop the electromagnetic torque, the voltages applied to the stator windings must be supplied as the functions of the rotor angular displacement. Therefore, the Hall-effect sensors must be used, or the so-called sensorless controllers (the rotor position is observed or estimated using the directly measured variables) must be designed and implemented using ICs. This brief discussion illustrates a wide spectrum of fundamental problems involved in the design of integrated microtransducers with controlling and signal processing ICs.

Conclusions

The critical focus themes in MEMS development and implementation are rapid synthesis, design, and prototyping through synergetic multi-disciplinary system-level research in electromechanics. In particular, MEMS devising, modeling, simulation, analysis, design and optimization, which is relevant to cognitive study, classification, and synthesis must be performed. As microtransducers and MEMS are devised, the fabrication techniques and processes are developed and carried out. Devising microtransducers is the closed evolutionary process to study possible system-level evolutions based upon synergetic integration of microscale structures and devices in the unified functional core. The ability to devise and optimize microtransducers to a large extent depends on the validity and integrity of mathematical models. Therefore, mathematical models for different microtransducers were derived and analyzed. It is documented that microtransducer modeling, analysis, simulation, and design must be based on reliable mathematical models which integrate nonlinear electromagnetic features. It is important to emphasize that the secondary phenomena and effects, usually neglected in conventional miniscale electromechanical motion devices (modeled using lamped-parameter models and analyzed using finite element analysis techniques) cannot be ignored. The fabrication processes were described to make high-performance microtransducers.

References

1. Campbell, S. A., *The Science and Engineering of Microelectronic Fabrication,* Oxford University Press, New York, 2001.
2. Lyshevski, S. E., *Nano- and Micro-Electromechanical Systems: Fundamental of Micro- and Nano-Engineering,* CRC Press, Boca Raton, FL, 2000.
3. Lyshevski, S. E., *MEMS and NEMS: Systems, Devices, and Structures,* CRC Press, Boca Raton, FL, 2001.
4. Madou, M., *Fundamentals of Microfabrication,* CRC Press, Boca Raton, FL, 1997.
5. Kim, Y.-J. and Allen, M. G., "Surface micromachined solenoid inductors for high frequency applications," *IEEE Trans. Components, Packaging, and Manufacturing Technology,* part C, vol. 21, no. 1, pp. 26–33, 1998.
6. Park, J. Y. and Allen, M. G., "Integrated electroplated micromachined magnetic devices using low temperature fabrication processes," *IEEE Trans. Electronics Packaging Manufacturing,* vol. 23, no. 1, pp. 48–55, 2000.

7. Sadler, D. J., Liakopoulos, T. M., and Ahn, C. H., "A universal electromagnetic microactuator using magnetic interconnection concepts," *Journal Microelectromechanical Systems,* vol. 9, no. 4, pp. 460–468, 2000.
8. Lyshevski, S. E., *Electromechanical Systems, Electric Machines, and Applied Mechatronics,* CRC Press, Boca Raton, FL, 1999.
9. Frazier, A. B. and Allen, M. G., "Uses of electroplated aluminum for the development of microstructures and micromachining processes," *Journal Microelectromechanical Systems,* vol. 6, no. 2, pp. 91–98, 1997.
10. Guckel, H., Christenson, T. R., Skrobis, K. J., Klein, J., and Karnowsky, M., "Design and testing of planar magnetic micromotors fabricated by deep x-ray lithography and electroplating," *Technical Digest of International Conference on Solid-State Sensors and Actuators, Transducers 93,* Yokohama, Japan, pp. 60–64, 1993.
11. Taylor, W. P., Schneider, M., Baltes, H., and Allen, M. G., "Electroplated soft magnetic materials for microsensors and microactuators," *Proc. Conf. Solid-State Sensors and Actuators, Transducers 97,* Chicago, IL, pp. 1445–1448, 1997.
12. Lagorce, L. K., Brand, O., and Allen, M. G., "Magnetic microactuators based on polymer magnets," *Journal Microelectromechanical Systems,* vol. 8, no. 1, pp. 2–9, 1999.
13. Smith, D. O., "Static and dynamic behavior in thin permalloy films," *Journal of Applied Physics,* vol. 29, no. 2, pp. 264–273, 1958.
14. Suss, D., Schreft, T., and Fidler, J., "Micromagnetics simulation of high energy density permanent magnets," *IEEE Trans. Magnetics,* vol. 36, no. 5, pp. 3282–3284, 2000.
15. Judy, J. W. and Muller, R. S., "Magnetically actuated, addressable microstructures," *Journal Microelectromechanical Systems,* vol. 6, no. 3, pp. 249–256, 1997.
16. Yi, Y. W. and Liu, C., "Magnetic actuation of hinged microstructures," *Journal Microelectromechanical Systems,* vol. 8, no. 1, pp. 10–17, 1999.
17. Gere, J. M. and Timoshenko, S. P., *Mechanics of Materials,* PWS Press, 1997.
18. Groom, N. J. and Britcher, C. P., "A description of a laboratory model magnetic suspension test fixture with large angular capability," *Proc. Conf. Control Applications, NASA Technical Paper – 1997,* vol. 1, pp. 454–459, 1992.
19. Ahn, C. H., Kim, Y. J., and Allen, M. G., "A planar variable reluctance magnetic micromotor with fully integrated stator and coils," *Journal Microelectromechanical Systems,* vol. 2, no. 4, pp. 165–173, 1993.
20. O'Sullivan, E. J., Cooper, E. I., Romankiw, L. T., Kwietniak, K. T., Trouilloud, P. L., Horkans, J., Jahnes, C. V., Babich, I. V., Krongelb, S., Hegde, S. G., Tornello, J. A., LaBianca, N. C., Cotte, J. M., and Chainer, T. J., "Integrated, variable-reluctance magnetic minimotor," *IBM Journal Research and Development,* vol. 42, no. 5, 1998.
21. Yasseen, A. A., Wu, C. H., Zorman, C. A., and Mehregany, M., "Fabrication and testing of surface micromachined polycrystalline SiC micromotors," *IEEE Trans. Electron Device Letters,* vol. 21, no. 4, pp. 164–166, 2000.
22. Yasseen, A. A., Zorman, C. A., and Mehregany, M., "Surface micromachining of polycrystalline silicon carbide films microfabricated molds of SiO and polysilicon," *Journal Microelectromechanical Systems,* vol. 8, no. 1, pp. 237–242, 1999.

IV

Systems and Controls

21
The Role of Controls in Mechatronics

Job van Amerongen
University of Twente

21.1 Introduction

"Mechatronic design deals with the integrated and optimal design of a mechanical system and its embedded control system." This definition implies that the mechanical system is enhanced with electronic components in order to achieve a better performance, a more flexible system, or just reduce the cost of the system. In many cases the electronics are present in the form of a computer-based embedded (control) system. This does not imply that every controlled mechanical system is a mechatronic system because in many cases the control is just an add-on to the mechanical system in a sequential design procedure. A real mechatronics approach requires that an optimal choice be made with respect to the realization of the design specifications in the different domains. In control engineering the design of an optimal control system is well understood and for linear systems standard methods exist. The optimization problem is formulated as: given a process to be controlled, and given a performance index (cost function), find optimal controller parameters such that the cost function is minimized. With a state feedback controller and a quadratic cost function, solutions for the optimal controller gains can be found with standard controller design software, such as Matlab[1] (Fig. 21.1).

Mechatronic design on the contrary requires that not only the controller be optimized. It requires optimization of the system as a whole. In the ideal case all the components in the system: the process itself, the controller, as well as the sensors and actuators, should be optimized simultaneously (Fig. 21.2).

In general this is not feasible. The problem is ill defined and has to be split into smaller problems that can be optimized separately. Later on the partial solutions have to be combined and the performance of the complete system has to be evaluated. After eventually readjusting some parts of the system this leads to a sub-optimal solution.

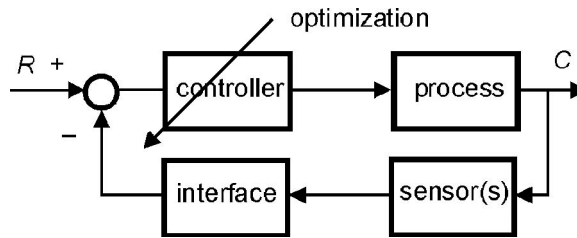



FIGURE 21.1 Optimization of the controller.

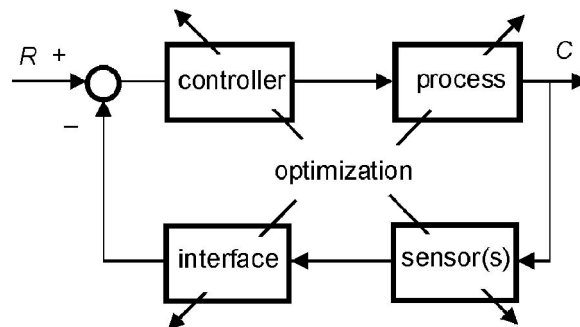



FIGURE 21.2 Optimization of the all system components simultaneously.

In the initial conceptual design phase it has to be decided which problems should be solved mechanically and which problems electronically. In this stage decisions about the dominant mechanical properties have to be made, yielding a simple model that can be used for controller design. Also a rough idea about the necessary sensors, actuators, and interfaces has to be available in this stage. When the different partial designs are worked out in some detail, information about these designs can be used for evaluation of the complete system and be exchanged for a more realistic and detailed design of the different parts.

Although the word mechatronics is new, mechatronic products have been available for some time. In fact, all electronically controlled mechanical systems are based on the idea of improving the product by adding features realized in another domain. Good mechatronic designs are based on a *real systems approach*. But mostly, control engineers are confronted with a design in which major parameters are already fixed, often based on static or economic considerations. This prohibits optimization of the system as a whole, even when optimal control is applied.

In the last days of gramophones, the more sophisticated designs used tacho feedback in combination with a light turntable to achieve a constant number of revolutions. But a really new design was the compact disc player. Instead of keeping the number of revolutions of the disc constant, it aims for a constant speed of the head along the tracks of the disc. This means that the disc rotates slower when tracks with a greater diameter are read. The bits read from the CD are buffered electronically in a buffer that sends its information to the DA converter, controlled by a quartz crystal. This enables the realization of a very constant bit rate and eliminates all audible speed fluctuations. Such a performance could never be obtained from a pure mechanical device only, even if it were equipped with a good speed control system. In fact, the control loop for the disc speed does not need to have very strict specifications. It should only prevent overflow or underflow of the buffer. The high accuracy is obtained in an open loop mode, steered by a quartz crystal (Fig. 21.3).

The flexibility introduced by the combination of precision mechanics and electronic control has allowed the development of CD-ROM players, running at speeds more than 50 times faster than the original audio CDs. A new way of thinking was necessary to come to such a new solution. On the other hand, the CD player is still a sophisticated piece of precision mechanics. No solid-state electronic memory

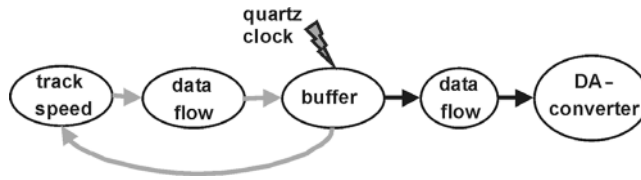



FIGURE 21.3 Combination of closed-loop and open-loop control in a CD player.

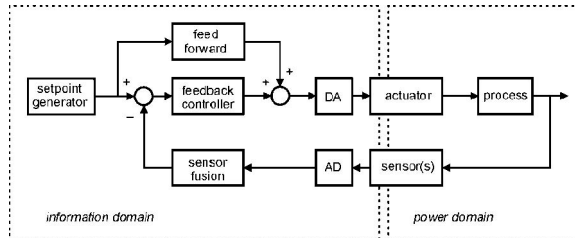



FIGURE 21.4 Mechatronic system.

device can compete yet economically with the opto-mechanical storage capabilities of the CD and its successor the DVD. But this may change rapidly.

21.2 Key Elements of Controlled Mechatronic Systems

A mechatronic system consists by definition of a mechanical part that has to perform certain motions and an electronic part (in many cases an embedded computer system) that adds intelligence to the system. In the mechanical part of the system power plays a major role. This in contrast to the electronic part of the system where information processing is the main issue. Sensors convert the mechanical motions into electrical signals where only the information content is important or even into pure information in the form of numbers (if necessary, through an AD converter). Power amplifiers convert signals into modulated power. In most cases the power supply is electrical, but other sources such as hydraulic and pneumatic power supplies are possible as well. A controlled mechanical motion system thus typically consists of a mechanical construction, one or more actuators to generate the desired motions, and a controller that steers the actuators based on feed-forward and sensor-based feedback control (Fig. 21.4).

21.3 Integrated Modeling, Design and Control Implementation

Modeling

During the design of mechatronic systems it is important that changes in the construction and the controller be evaluated simultaneously. Although a proper controller enables building a cheaper construction, a badly designed mechanical system will never be able to give a good performance by adding a sophisticated controller. Therefore, it is important that during an early stage of the design a proper choice can be made with respect to the mechanical properties needed to achieve a good performance of the controlled system. On the other hand, knowledge about the abilities of the controller to compensate for mechanic imperfections may enable that a cheaper mechanical construction be built. This requires that in an early stage of the design a simple model is available that reveals the performance limiting factors of the system. Still there is a gap between modeling and simulation software used for evaluation of mechanical constructions and software used for controller design. Mechanical engineers are used to

finite element packages to examine the dynamic properties of mechanical constructions. It is only after reduction to low-order models (modal analysis) that these models can be used for controller design. On the other hand, typical control-engineering software does not directly support the mechatronic design process either; in the modeling process the commonly used transfer functions and state space descriptions often have lost the relation with the physical parameters of the mechanical construction. Tools are required that allow modeling of mechanical systems in a way that the dominant physical parameters (like mass and dominant stiffness) are preserved in the model and simultaneously provide an interface to the controller design and simulation tools control engineers are used to (Coelingh;[2] Coelingh, De Vries, and Van Amerongen[3]).

Simulation is an important tool to evaluate the design of mechatronic systems. Most simulation programs like Simulink[1] use block diagram representations and do not support physical modeling in a way that direct tuning of the physical parameters of the mechanical construction and those of the controller is possible as required in the design of mechatronic systems. Recently, programs that allow physical modeling in *various physical domains* became available. They use an object-oriented approach that allows hierarchical modeling and reuse of models. The order of computation is only fixed after combining the subsystems. Examples of these programs are 20-sim,[4] described by Broenink[5] as CAMAS and Dymola.[6]

In this section the modeling and simulation program 20-sim (pronounced Twente Sim) will be used to illustrate the simultaneous design of construction and controller in a mechatronic system. 20-sim supports object-oriented modeling. Power and signal ports to and from the outside world determine each object (Weustink, De Vries, and Breedveld[7]). Inside the object there can be other objects or, on the lowest level, equations. Various *realizations* of an object can contain different or more detailed descriptions as long as the interface (number and type of ports) is identical. Modeling can start by a simple interconnection of (empty) submodels. Later they can be filled with realistic descriptions with various degrees of complexity. De Vries[8] refers to this as *polymorphic* modeling. Submodels can be constructed from other submodels in hierarchical structures. Proper physical modeling is achieved by coupling the submodels by means of the *flow of energy*, rather than by *signals* such as voltage, current, force, and speed. This way of modeling is well suited for mechatronics system design. It will be illustrated with an example. We want to consider the design of a simple servo system, considering the use of a voltage source, a DC motor, and a mechanical load driven through a transmission (Fig. 21.5).

The transmission is disregarded for the time being. The belt is considered as infinitely stiff and the transformation ratio is taken care of by changing the motor constant. If a power amplifier driven by a signal generator describes the voltage source, we can draw the iconic diagram of Fig. 21.6. At this stage the different *components* in this model are still empty. But all components have electrical and/or mechanical "ports." With the proper interfaces (ports) defined, the components can be connected to each other.

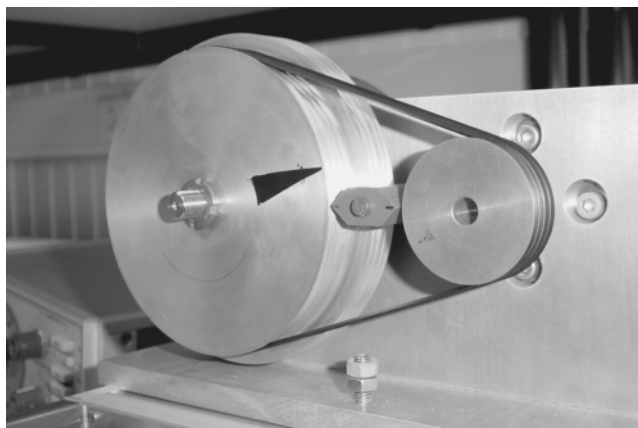



FIGURE 21.5 Simple DC-servo system.

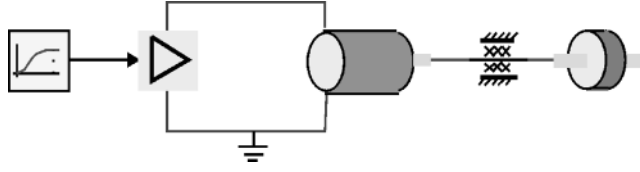



FIGURE 21.6 Iconic diagram of the simple servo system.

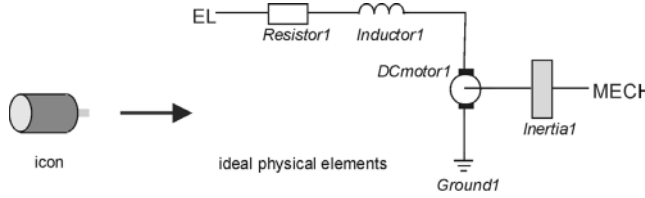



FIGURE 21.7 Icon of the motor expanded to ideal physical elements.

In the next step we can detail the description of the DC motor. One solution could be the description given in Fig. 21.7. The motor is now described by a number of *ideal physical elements*, each representing a basic physical relation. The motor has an electrical (EL) as well as a mechanical port (MECH).

Each of the *elements* in this figure can be described as an element with an electrical and/or mechanical port. The idea of ports is made more explicit in so-called bond graphs.[9–12] For the electrical elements these are the voltage difference over the element and the current through the element. For the mechanical elements these are the torque and the (angular) velocity. The products of these conjugated variables ($P = ui$ or $P = T\omega$) represent power.

If we go down a step further into the hierarchy, we arrive at the level of equations. For instance, an electrical resistor can be described by the equation:

$$p \cdot u = R^* p \cdot i \tag{21.1}$$

where the variables $p \cdot u$ and $p \cdot i$ indicate the conjugated variables u and i of the electrical port p. Note that this is an equation and not an assignment statement. It could have been written equally well in the form:

$$p \cdot i = 1/R * p \cdot u \tag{21.2}$$

In a similar way the inductance can be described by the equations:

$$p \cdot u = L^* \text{ddt}(p \cdot i) \text{or} \qquad (p \cdot i) = 1/L * \text{int}(p \cdot u) \tag{21.3}$$

where ddt($p \cdot i$) denotes di/dt and int($p \cdot u$) denotes $\int u\, dt$. In case of an R-element there is no preference for one of the two forms. For the I-element the integral form is preferred in the simulations. 20-sim determines the preferred causal form and derives the equations automatically.

The energy flow or *power P* is the product of *two conjugated signals*, called effort (e) and flow (f):

$$P = ef \tag{21.4}$$

Examples of this expression in the mechanical and electrical domain are

$$P = Fv \qquad \text{or} \qquad P = T\omega \tag{21.5}$$

$$P = ui \tag{21.6}$$

where F is force, v is velocity, T is torque, ω is angular velocity, u is voltage, and i is current.

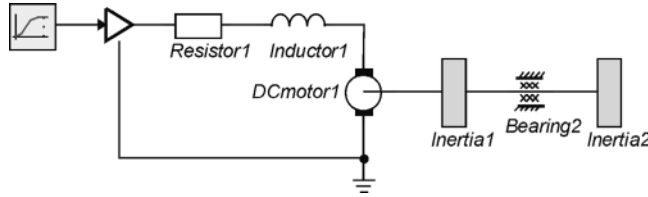



FIGURE 21.8 Complete model in the form of ideal physical elements.

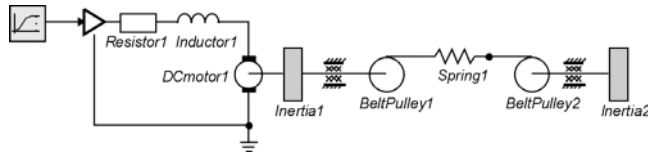



FIGURE 21.9 Model extended with transmission.

When we expand the complete Fig. 21.6 we obtain Fig. 21.8. When this model is processed a message pops up that indicates that inertia 2 has a dependent state. The two inertias in this model always have the same speed, and therefore, they are dependent. They cannot have independent initial conditions. The message indicates that this element can only be written in derivative form:

$$T = J\ d\omega/dt \tag{21.7}$$

There are several ways to deal with this problem.

1. The two inertias can be combined into one inertia (the program will do this automatically). A message pops up that the dependency of the two inertias has been solved symbolically.
2. Dealing with the derivative causality by means of an implicit integration algorithm.
3. The transmission can be added, including some flexibility in the belt.

If the flexibility is negligible, solution 1 leads to the simplest model. On the other hand, the warning raises the question whether the flexibility of the belt can be disregarded indeed. If not, the model has to be extended with a spring element. It should be noted that this should not be done for numerical reasons only. If the transmission were very stiff, this would result in high-frequency dynamics and lead to unnecessary slow simulations. On the other hand, if the flexibility is important, as it is in this system, the warning draws the designer's attention to the fact that the model may be oversimplified. In Fig. 21.9 the transmission, including a spring element, has been added. Processing of this model does not produce any warnings.

This example illustrates how modern software can help to come up with a model that has the complexity that is needed for a particular problem. Physical models, in the form of an iconic diagram, based on connecting elements by means of power ports, may help in this modeling process. The user can select the preferred view, whether this is a bond graph, an iconic diagram with ideal physical element, or a view using higher lever submodels, like in Fig. 21.6. In the next section it will be shown how to use this model for the design of controllers.

Control System Design Methodologies

Many processes can be reasonably well controlled by means of PID controllers. This is due to the fact that these processes can be more or less accurately described by means of a second-order model. Tuning rules, like those of Ziegler Nichols, enable less experienced people to tune such controllers. Relatively simple

models can also describe many mechatronic systems. A mechatronic system mostly consists of an actuator, some form of transmission, and a load. A fourth-order model can properly describe such a system. The performance-limiting factor in these systems is the resonance frequency. A combination of position and tacho feedback (basically a PD controller) can be applied here as well. But due to the resonant poles proper selection of the signals to be used in the feedback is essential. Efforts have been made (Groenhuis;[13] Coelingh;[2] Coelingh, De Vries, and Van Amerongen[3]) to derive recipes for tuning such systems, in addition to selecting the proper feedback signals. Computer support tools are essential to enable less experienced designers to use these recipes (Van Amerongen, Coelingh, and De Vries[14]). Coelingh[2] and Coelingh, De Vries, and Van Amerongen[3] describe a structural design method for mechatronic systems. The method starts with reducing the conceptual design to a fourth-order model that represents the dominant properties of the system in terms of the total mass to be moved and the dominant stiffness. This model still has physical meaningful parameters. In this model appropriate sensors are chosen, as well as a path generator. In the conceptual design phase a simple controller is developed and mechanical properties are changed, if necessary. Then a more detailed design phase follows where also parameter uncertainties are taken into account.

Servo System Design

Here we will consider some simple aspects of the design of a servo system in order to illustrate the advantage of the use of physical models and to illustrate the need for an integrated design approach. We consider the model discussed before, a load driven by an electric motor, through a flexible transmission. The iconic diagram of this model was given in Fig. 21.9. In this example a current amplifier has replaced the voltage amplifier allowing the removal of the electrical resistor and the inductance. In the step responses of Fig. 21.10 the resonance due to the flexible transmission is clearly visible.

From the equations used for the simulation, 20-sim can automatically derive a model in a form suitable for controller design, such as a state-space description, a transfer function, or poles and zeros. An interface is provided to Matlab[1] enabling, for instance, to use Matlab algorithms to compute the gains of advanced controllers like an LQR (optimal state feedback) or LQG controller (with a Kalman filter for state estimation and optimal state feedback). The diagram of the process together with an LQG controller is given in Fig. 21.11 and some responses in Fig. 21.12.

A properly designed P(I)D controller is able to perform almost similarly, especially when the amount of noise is small. A first attempt could be to use only measurements of the load angle and load speed.

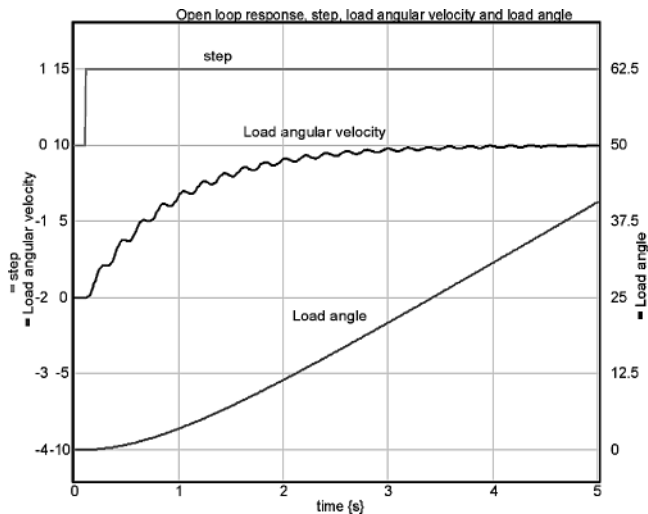



FIGURE 21.10 Open loop responses.

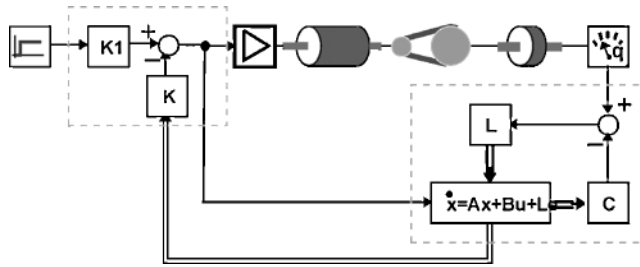



FIGURE 21.11 Process with Kalman filter and state feedback.

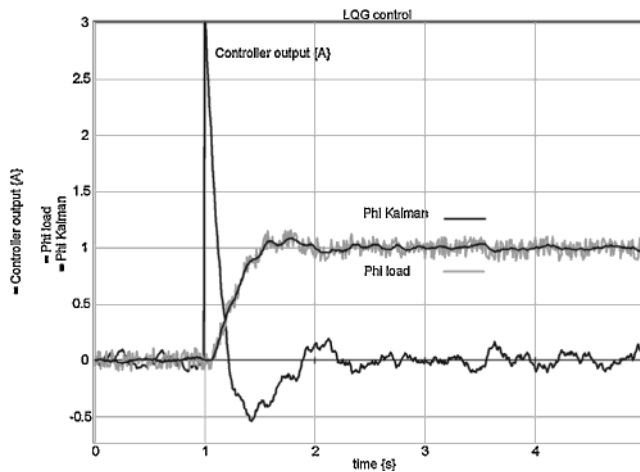



FIGURE 21.12 Response of the LQG-controlled system.

This attempt fails, because feedback of the load speed leads almost immediately to an unstable system as can be seen from the root locus for variations in the gain of the velocity feedback. From the responses of Fig. 21.10, 20-sim can easily determine the transfer function between the motor current and the load speed and plot the root locus (Fig. 21.13).

Figure 21.14 clearly shows that even a small amount of velocity feedback will lead to an unstable system. It is well known that feedback of the motor speed is a better solution. Using again the model of Figs. 21.9 and 21.10 to determine the transfer from input current to motor speed yields the root locus of Fig. 21.15.

Complex zeros now accompany the complex poles and because they are close together their influence on the response will be almost negligible. The branch of the root locus on the real axis now shows the desired behavior: moving the dominant pole to the left in the *s*-plane. Combining the feedback of the motor speed with feedback of the load angle yields the PD-controller structure of Fig. 21.15 and the responses of Fig. 21.16. Except for the noise there is not much difference in the responses of the system with the Kalman filter, although the PD-controlled system is simpler. The observations made here are generally applicable. A system with two resonant (complex) poles and no zeros, such as in Fig. 21.13, is difficult to control by means of a simple controller. If complex zeros accompany the resonant poles with an imaginary part smaller than that of the poles, stable control is easily achieved. In the frequency domain this is seen as an anti-resonance, followed by a resonance (type *AR*). On the contrary a type *RA* system, where the resonance frequency is lower than the anti-resonance frequency (the imaginary part of the poles is smaller than that of the zeros), is just as difficult to control as in the case of only resonant poles. The existence and location of resonant zeros is completely determined by the (geometrical) location of

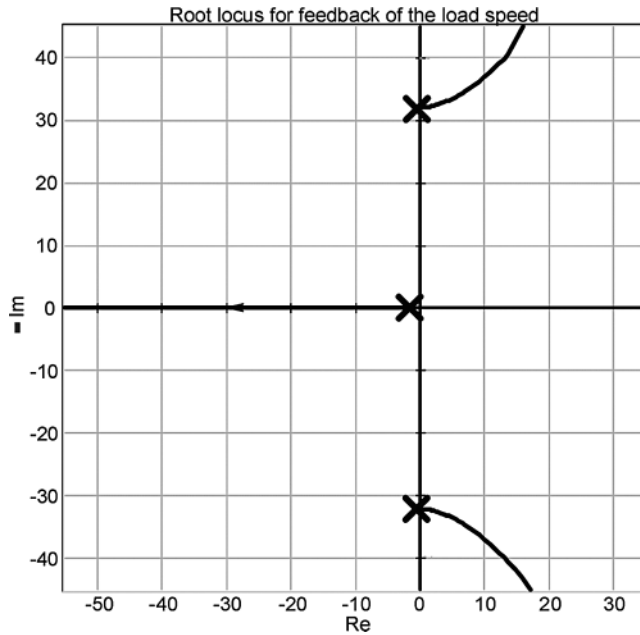



FIGURE 21.13 Root locus for velocity feedback of load axis.

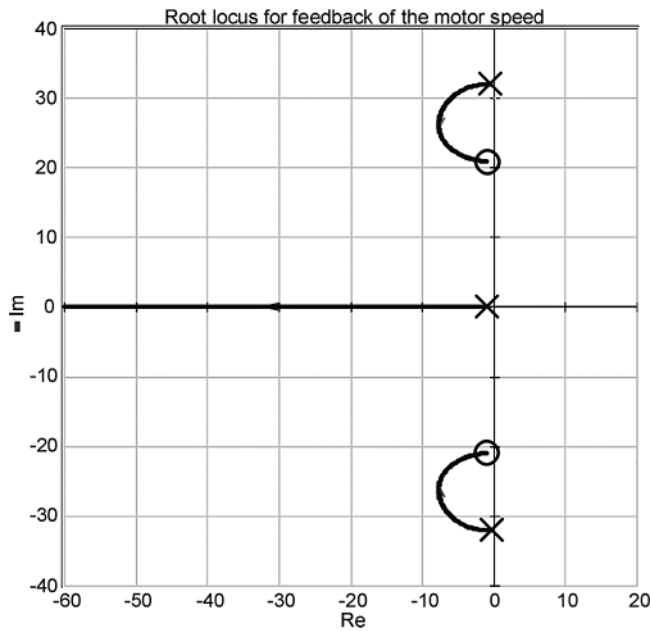



FIGURE 21.14 Root locus for velocity feedback of motor axis.

the sensors in the mechanical system. A careful choice of these sensor locations is therefore crucial for the successful application of a controller. It should be noted that using a properly designed set-point generator could prevent resonance, as seen in Fig. 21.10. The set point generator should not excite the resonance frequencies, for instance, by using a low pass filter with bandwidth lower than the resonance frequencies. However, such a set-point generator does not solve the above-mentioned stability problems.

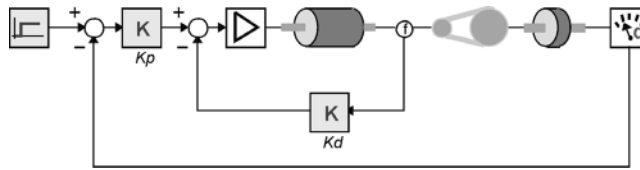



FIGURE 21.15 Servo system with PD-controller.

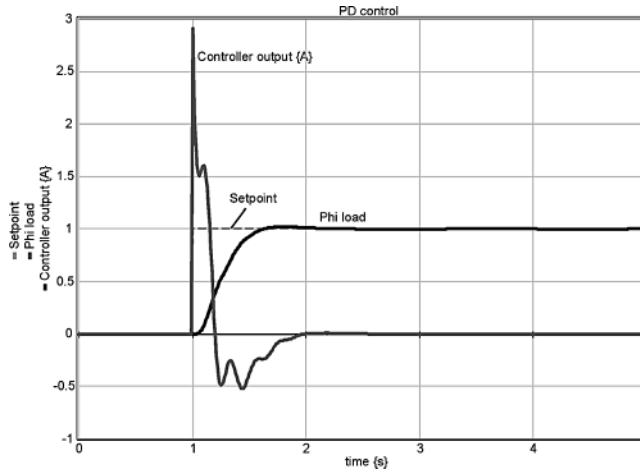



FIGURE 21.16 Responses of the system of Fig. 21.16.

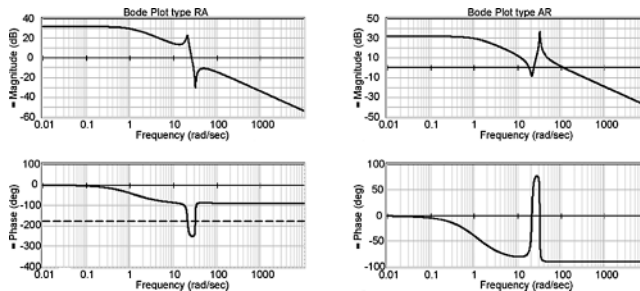



FIGURE 21.17 Bode plots of type *RA* and *AR* systems.

Design of a Mobile Robot

A typical example of the early design procedure is the conceptual design of a mobile assembly robot. Already in a very early stage of the design conflicting demands have to be resolved. Such a robot should be able to collect parts all around a production facility and do the assembly while driving. Because a high accuracy is required between the gripper of the robot and the surface where the parts are located, it is important that floor irregularities and vibration modes of the structure do not prevent proper assembly. On the other hand the path controller, partly based on dead reckoning (i.e., measuring of the wheel speed and orientation), requires that the wheels be very stiff. Damping of disturbances has to be realized by another means of suspension. This has led to the concept of an upper frame and a lower frame, connected by means of springs (Fig. 21.18).

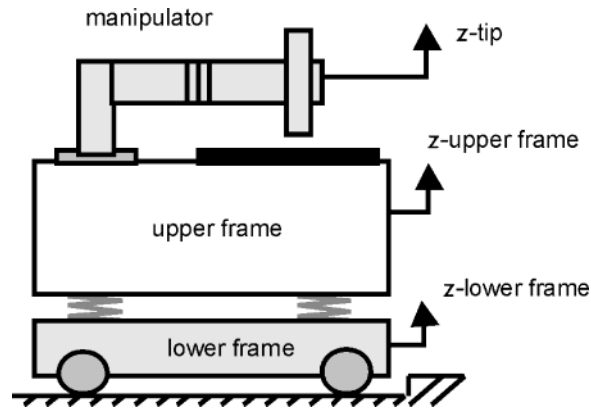



FIGURE 21.18 Conceptual design of the mobile robot.

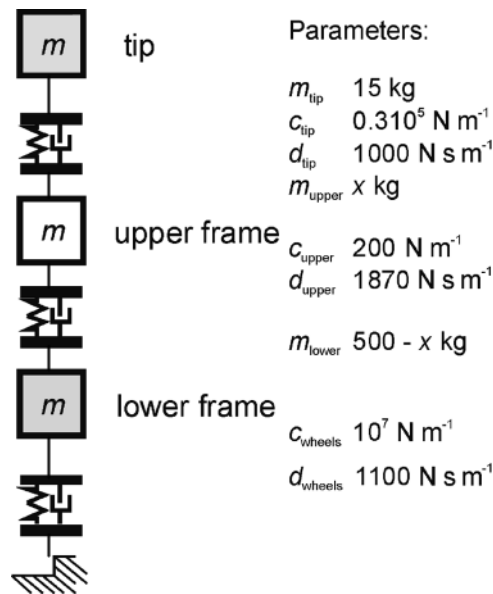



FIGURE 21.19 Simple model with ideal physical elements to compute the error e_{tip}.

The robot can be mounted at the upper frame and should have sufficient bandwidth such that the position error ($e_{tip} = z_{tip} - z_{upper\ frame}$) between the tip of the robot (z_{tip}) and the upper frame ($z_{upper\ frame}$) is small enough.

The next step is to derive a simple model, in order to have some parameters for the weight distribution and the stiffness and damping of the springs. In the model of Fig. 21.22 the robot is confronted with a bump in the floor at a speed of 1 m/s.

Based upon the payload—mainly the weight of the batteries—the total mass of the vehicle was estimated to be 500 kg. Stiffness and damping of the wheels follow from the demands for the accuracy of the position estimation. The mass and bandwidth of the controlled manipulator were already known from other studies, yielding the effective stiffness and damping for the robot tip. When also initial estimates of the stiffness and damping of the springs between the upper and lower frame are made, the only parameter to be varied is the weight distribution between the upper frame and lower frame. By using the optimization feature of 20-sim, the optimal weight distribution can easily be found. In order to minimize the error between the tip of the robot and the upper frame (Fig. 21.19), the weight has to be placed as much as possible in the

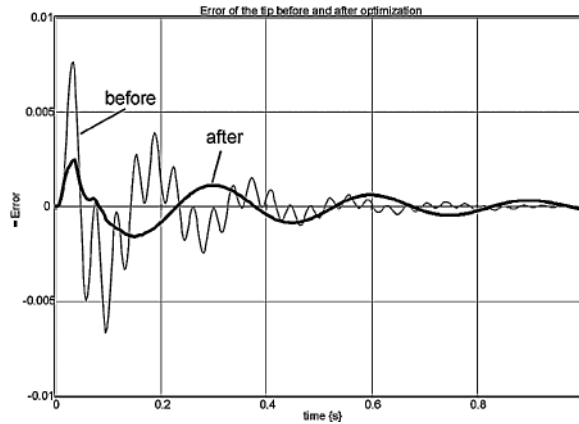



FIGURE 21.20 Error of the tip before and after optimization of the weight distribution between upper and lower frame.

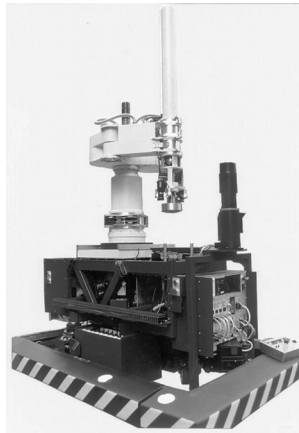



FIGURE 21.21 The mobile robot (MART) after completion.

upper frame (Fig. 21.20). This example illustrates how the mechanical configuration of the system is determined by the requirements for good path control and accurate control of the assembly task.

A next step could be to optimize the properties of the suspension between upper and lower frame. This will further improve the error. This decision made in a very early stage of the design directed other design decisions. After completion of the project it appeared that the different parameters of the final construction were close to these early estimates (Fig. 21.21).

21.4 Modern Examples of Mechatronic Systems in Action

A few examples have already been treated in the previous sections. In this section two more examples will be given.

Rudder Roll Stabilization of Ships

Nowadays most ships use an autopilot to control the heading of the ship. A rudder is the most commonly used actuator. Some ships, like ferries and naval ships, need also roll stabilization. This can be achieved passively by means of two connected tanks filled with water that generate stabilizing forces that should

be in counter phase with the forces of the waves. In order to make the system effective for varying frequencies of the waves, the water flow between the two tanks should be controlled. For fast ships mostly stabilizing fins are used. These are a kind of actively controlled "wings" that generate the moments needed to counteract the moments of the waves. The fins not only influence the roll motions but also have influence on the heading. On the other hand, the rudder not only influences the heading but also induces roll. In control engineering terms this leads to a multivariable system that requires a multivariable controller design for optimum performance. In practice such a multivariable system is seldom seen and two separate control systems are used.

Another approach is to use only one of the actuators (rudder or fins) to achieve course control and roll reduction. Because the frequencies of the roll motions are outside the bandwidth of the course-control system this is possible. The rudder is most suited as actuator. An additional advantage for naval ships is that removing the fins will reduce the underwater noise of the vessel.

Redesigning the course controller in order to stabilize the roll as well, demonstrates the feasibility of this approach, but also makes clear that the "process"—the ship—should be modified. The most important modification is needed for the steering machine. The maximum speed of the steering machine appears to be the limiting factor for such a system (it should increase from the commonly used values of 3–7∞/s to 20–25∞/s). By means of dynamic simulations the demands for the steering machine can be found in terms of the maximum speed of the steering machine and the maximum time constant that is allowed for reaching this speed. This requires reengineering of the hydraulic steering machine. A step further would be to consider also changes in the shape of the ship, in order to optimize the parameters that determine the effectiveness of the rudder roll stabilization system.

In order to decide whether this new solution is better, it should be evaluated whether the redesigned steering machine is less expensive than the original rudder and fin actuators. These design issues have to be solved in a very early stage of the design. Rudder roll stabilization has been successfully applied on naval as well as merchant marine ships (Van Amerongen, Van der Klugt, and Van Nauta Lemke[15]).

Compensation of Nonlinear Effects in a Linear Motor

Many mechanical systems suffer from nonlinear effects that limit the accuracy that can be achieved. Friction and cogging are two examples. A (linear) feedback controller can diminish the influence of nonlinearities, but complete compensation may be difficult. For systems that perform repetitive motions, an Iterative Learning Controller can help to further improve the performance (Arimoto;[16] De Vries, Velthuis, and Van Amerongen[17]). The basic idea is explained in Fig. 21.22.

When only the feedback loop is present and under the assumption that there are no disturbances, the error signal and thus the controller signal U_C will be the same for each repetitive motion. It is obvious that the accuracy can be improved when in the next motion the controller signal from the former cycle is used as a feed-forward signal, U_F. The feedback will generate a signal that further compensates for the remaining error by updating the feed-forward signal U_F with the formula

$$U_F^{k+1} = U_F^k + LE^k \tag{21.8}$$

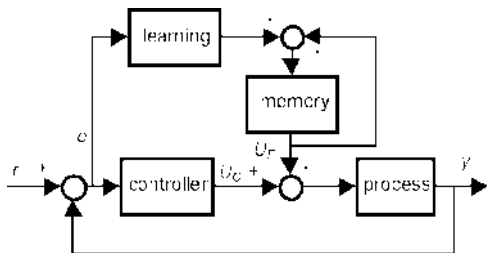



FIGURE 21.22 Principle of Iterative Learning Control.

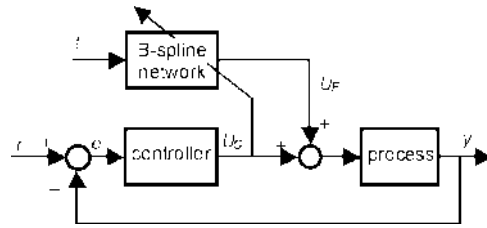



FIGURE 21.23 Learning feed-forward controller for repetitive motions.

where L is the transfer function of the learning filter. The superscript k denotes the kth repetitive motion. The signal U_F should converge to a feed-forward signal that compensates for all repetitive errors. An example of a situation where such errors are present is, for instance, a CD player that has to compensate for the eccentricity of the disk.

A variation on this idea and even more straightforward is the learning feed-forward controller (LFFC) setup of Fig. 21.23. When the feed-forward signal would be perfect, the output of the controller would be zero. This implies that this output can be used as a training signal for a neural network. An adaptive B-spline network enables learning of complex nonlinear characteristics. Also support vector machines have been used to implement the learning feed forward (De Kruif and De Vries[18]). The input of the B-spline network is the time t. It is reset each time a new motion starts. This is called a time-indexed LFFC. Instead of the time, also the reference signal and its derivatives—obtained from a path generator—could be used as index for the network (path-indexed LFFC). The advantage of this structure is that after proper training the LFFC can successfully be used for nonrepetitive motions as well. Velthuis[19] has given a stability analysis for time-indexed as well as path-indexed LFFC-controllers. The stability analysis is relatively easy for the time-indexed case. For the path-indexed case it is more complex and some heuristics are required to guarantee a stable system. The main issue is that the number of B-splines should not be too large. On the other hand a sufficiently dense B-spline distribution is desired for an accurate approximation of the nonlinear process. LFFC has successfully been applied to compensate for cogging in an industrial Linear Motor (Otten et al.[20]) and for compensation of (Coulomb) friction of a linear motor used in a flight simulator (Velthuis[19]). It has also been applied to the tracking control of the mobile robot described in the section Design of a Mobile Robot (Starrenburg et al.[21]).

The application to cogging compensation of a linear motor will be described in a little bit more detail. Such a motor is a commonly used element in assembly machines. Even with the best magnets and accurate assembly the error could not be made smaller than 100 μ, with a PID controller in combination with nonlearning feed-forward control. The design goal was to improve the maximally achievable accuracy from 100 μ to less than 10 μ. Figure 21.24 shows a picture of a linear motor.

According to the structure of Fig. 21.23 the linear motor is controlled by means of a PID controller, while a B-spline neural network is present to learn the inverse motor model, including the nonlinearity due to cogging. Cogging occurs in DC motors with permanent magnets. It causes more or less sinusoidal shaped forces that depend on the position of the translator with respect to the stator. If these forces really had a sinusoidal shape, they would be easy to compensate for by means of a feed-forward compensator. However, this would require magnets with completely similar magnetic properties and very accurate spacing of the magnets. An alternative is to design a controller that learns the disturbance pattern and compensates it by means of a learning feed-forward compensator. An additional advantage is that such a system can also be used to compensate for other nonlinear effects, such as friction. This has also been demonstrated in a part of a flight simulator (a control stick) where friction forces spoil the feeling of a realistic simulation especially at almost zero speed. Figure 21.25 shows that learning is almost completed after six training cycles.

Learning feed-forward control is an attractive method to compensate for nonlinearities that are present in mechatronic systems, such as cogging and friction. The use of B-spline neural networks results in fast convergence, relatively low computational effort, and a good generalizing ability. Because of recently obtained results with respect to the stability of such systems, robust control systems can be designed.

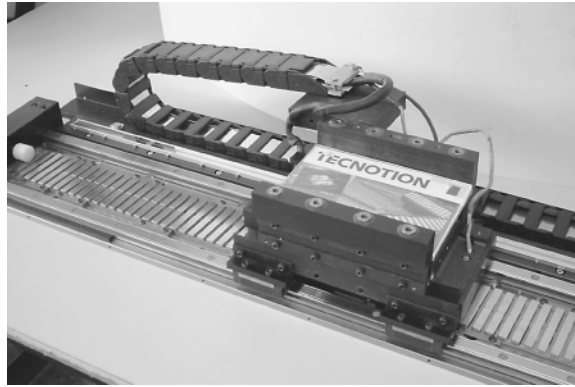



FIGURE 21.24 Linear motor. The magnets that cause the cogging are clearly visible.

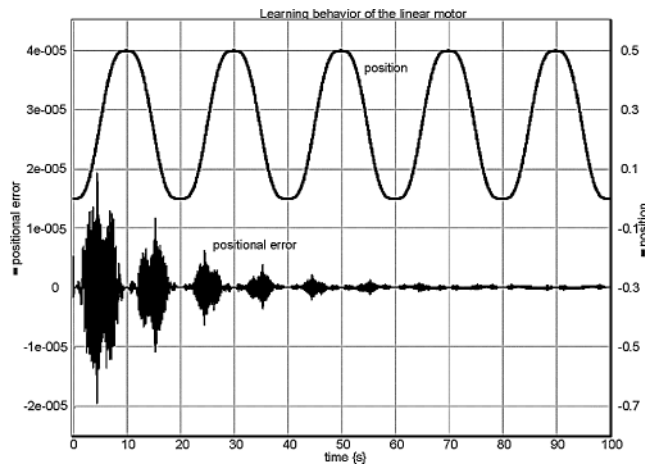



FIGURE 21.25 Position and error signal during learning of the LFFC.

A mechatronic view on this design problem raises the question whether it is possible to use the same techniques to build a less expensive linear motor, when maximum accuracy is not the main goal of the design. It has been demonstrated that a motor constructed with less expensive components and less demanding assembly specifications but with LFFC can compete well with the more expensive construction. The accuracy can typically be improved by a factor 10.

21.5 Special Requirements of Mechatronics that Differentiate from "Classic" Systems and Control Design

The main difference between "ordinary controller design" and mechatronic system design is that the latter deals with the design of the system as a whole. This approach can be considered as optimization of all components of the system simultaneously, although there are no algorithms to do this automatically. In practice the problem is often split into smaller problems that can be optimized. After integration of all the partial solutions a suboptimal system is achieved that can be further optimized by retuning the different parts, taking into account the already available intermediate design of the overall system. In order to achieve optimization of the system as a whole, it is desired that the mechanical part, where power plays a role, and the information processing part (the controller) can be simulated and adjusted simultaneously. This requires that mechanical parameters like masses and compliances be available in simulations of the

controlled system. Examples have been given of modeling and simulation with 20-sim that allows for such an approach.

Mechatronic designers should constantly be aware of the fact that solutions can be found in different domains. Not every mechanical deficiency can easily be solved by control. A good mechanical design may be easier and cheaper to achieve. On the other hand, a good controller may be able to achieve the desired performance much easier and cheaper than a complex mechanical construction. In some cases the combination can even achieve performances that would never have been possible without a mechatronic design.

The same holds for the design of sensors. Each sensor could be fitted with a filter to remove noise from the measurements. But if several sensors are being combined, sensor fusion in a Kalman filter algorithm will benefit from the availability of the raw data.

Communication between all the designers involved and transparency of the design decisions in the various domains are essential for the success of a true mechatronic design.

References

1. Matworks (2001). *The Mathworks: Developers of Matlab and Simulink,* www.mathworks.com
2. Coelingh, H.J., *Design Support for Motion Control Systems,* Ph.D. thesis, University of Twente, 2000, also www.rt.el.utwente.nl/clh/
3. Coelingh, H.J., de Vries, T.J.A., van Amerongen, J., Design support for motion control systems—application to the Philips fast component mounter, in *Mechatronics Forum 7th Int. Conf., Mechatronics 2000,* Atlanta, Ga, USA.
4. Controllab Products, *20-sim*, www.20sim.com
5. Broenink, J.F., *Computer-Aided Physical-Systems Modeling and Simulation: a Bond-Graph Approach,* Ph.D. thesis, University of Twente, 1990.
6. Dynasim, *Dymola,* www.dynasim.se/
7. Weustink, P.B.T., de Vries, T.J.A., Breedveld, P.C., *Object Oriented Modeling and Simulation of Mechatronic Systems with 20-sim 3.0,* in *Mechatronics 98,* J. Adolfson and J. Karlsén (Eds.), Elsevier Science, 1998.
8. De Vries, T.J.A., *Conceptual Design of Controlled Electro-Mechanical Systems,* Ph.D. thesis, University of Twente, 1994.
9. Breedveld, P.C., Fundamentals of bond graphs, in *IMACS Annals of Computing and Applied Mathematics, Vol. 3: Modelling and Simulation of Systems,* Basel, 1989, pp. 7–14.
10. Cellier, F.E., Elmqvist, H., Otter, M., Modeling from physical principles, in *The Control Handbook,* W.S. Levine (Ed.), CRC Press, pp. 99–108, 1996.
11. Gawthrop, P., Lorcan Smith, L., *Metamodelling: Bond Graphs and Dynamic Systems,* Prentice-Hall, NJ, 1996.
12. Van Amerongen, J., Modelling, simulation and controller design for mechatronic systems with 20-sim 3.0, in *Proc. 1st IFAC Conf. on Mechatronic Systems,* Darmstadt, Germany, September 2000, pp. 831–836.
13. Groenhuis, H., *A Design Tool for Electromechanical Servo Systems,* Ph.D. thesis, University of Twente, 1991.
14. Van Amerongen, J., Coelingh, H.J., de Vries, T.J.A., "Computer support for mechatronic control system design," *Robotics and Autonomous Systems,* vol. 30, no. 3, pp. 249–260, PII: SO921-8890 (99)00090-1, 2000.
15. Van Amerongen, J., van der Klugt, P.G.M., van Nauta Lemke, H.R., Rudder roll stabilization for ships, *Automatica,* vol. 26, no. 4, pp. 679–690.
16. Arimoto, S., A brief history of iterative learning control, in *Iterative Learning Control: Analysis, Design, Integration and Applications,* Kluwer Academic Publishers, pp. 3–7, 1988.
17. De Vries, T.J.A., Velthuis, W.J.R., van Amerongen, J., Learning feed-forward control: a survey and historical note, in *1st IFAC Conf. on Mechatronic Systems,* Darmstadt, Germany, September 2000, pp. 949–954.

18. De Kruif, Bas J., de Vries, T.J.A., On using a support vector machine in learning feed-forward control, in *Proc. 2001 IEEE/ASME Int. Conf. on Advanced Intelligent Mechatronics,* Como, Italy, 8–12 July, 2001.
19. Velthuis, W.J.*R., Learning Feed-Forward Control—Theory, Design and Applications,* Ph.D. thesis, University of Twente, 2000, also http://www.rt.el.utwente.nl/vts/
20. Otten, G., de Vries, T.J.A., van Amerongen, J., Rankers, A.M., Gaal, E., Linear motor motion control using a learning feedforward controller, *IEEE/ASME Transactions on Mechatronics,* vol. 2, no. 3, ISSN 1083-4435, pp. 179–187, 1997.
21. Starrenburg, J.G., van Luenen, W.T.C., Oelen, W., van Amerongen, J., Learning feed-forward controller for a mobile robot vehicle, *Control Engineering Practice,* vol. 4, no. 9, pp. 1221–1230, 1996.

22

The Role of Modeling in Mechatronics Design

Jeffrey A. Jalkio
University of St. Thomas

If mechatronics design is more than just the combination of electronic, software, and mechanical design, the additional feature must lie in the ability of the mechatronic designer to optimize a design solution across these disparate fields. This requires a sufficient understanding of each of these fields to determine which portions of an engineering problem are best solved in each of these domains given the current state of technology. In turn, this requires the ability to model the problem and potential solutions using techniques that are domain independent or at least permit easy comparison of solutions and tools from different domains.

For example, the optical inspection system shown in Fig. 22.1 depends on optical components in precise alignment, mechanical elements capable of precise motion, transducers for sensing and providing mechanical power, electrical systems to control motion and filter sensor signals, and software for image analysis and motion control. Only by dividing these tasks appropriately among electronics, mechanical components, and software can the system be optimized. This requires an understanding of all the system requirements and limitations as well as the capabilities of each component in the various domains. Modeling of requirements and systems is crucial in determining whether a proposed solution is acceptable as well as in documenting these determinations for future use. In this article we shall examine the varieties of models used at different points in the design process, the diverse roles of these models and their relative strengths and weaknesses in each of these roles, and finally the specific tradeoffs involved in choosing dynamic models for signals and systems analysis.

22.1 Modeling as Part of the Design Process

Models serve different purposes at different points in the design process; so to decide which modeling tools are most effectively employed in different phases we must examine the design process itself. Many descriptions of the design process are available that have been developed by researchers around the world.[1–3] Typically these descriptions serve to systematize the process to improve the productivity of

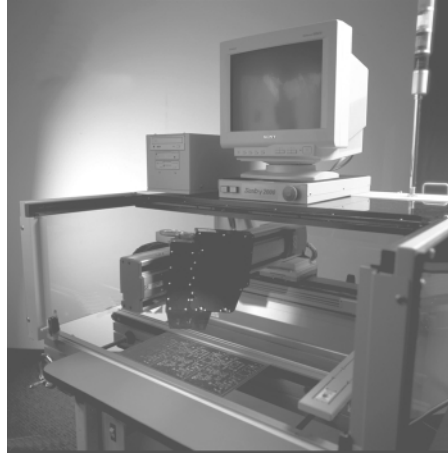



FIGURE 22.1 An optical inspection system for printed circuit boards. (Used by permission ” CyberOptics Corporation 2001, all rights reserved.)

designers or to describe techniques that provide improved product quality, lower cost, or other benefits. However, since our purpose is to examine the modeling needs of the design process, we can consider a simple model that distinguishes phases of the design process in terms of types of design activity rather than a more complex model that may be preferable for other purposes. For this purpose, we can consider a four-phase process consisting of requirements analysis, concept generation, analysis and selection, and detailed design. In the first phase of this process the designer focuses on analysis of the problem without considering possible solutions. In the second phase, conceptual solutions are generated with the hope that an acceptable solution can be found from these initial concepts via combination or modification of concepts or by variation of parameters present in one of the conceptual solutions. In the third phase, these concepts are evaluated and a design is chosen for implementation. The fourth phase consists of identifying design problems that need to be solved to implement the chosen concept and applying the design process to those smaller problems. We shall consider the activities of each of these phases in detail.

Phase 1

The requirements analysis phase consists in obtaining a sufficient understanding of the problem to be solved. The difficulty of this process varies with the scale of the problem, the designers' familiarity with the problem domain, the variability of market needs, and the presence of hidden requirements that are poorly articulated in the initial problem statement. Depending on the nature of the design problem, the requirements identified in this phase may be the needs of a single customer, the common needs of a group of potential customers identified via a market survey, or societal needs identified by government regulations. Most design problems include some combination of these as well as internal requirements such as design guidelines and company policies. The key objective of this phase is to obtain enough detail to know when a design has solved the problem satisfactorily. Models in this phase serve primarily as communication and documentation tools since the primary problem in this phase is the clear communication and documentation of the criteria for design success. Examples of models that aid in this process include specification listings, use case diagrams, sequence diagrams, and context diagrams. Many of these modeling tools have now been standardized as parts of the Unified Modeling Language (UML), which is becoming increasingly important as an analysis tool.[4]

Use case diagrams model the interactions between a system and its users at a very high level of abstraction in terms of purpose of the interaction. Figure 22.2 gives an example of a use case diagram used to document the various operations required of a network printer. The use case diagram helps us avoid overlooking important but rare use cases such as maintenance. It is important to note that use case diagrams do not

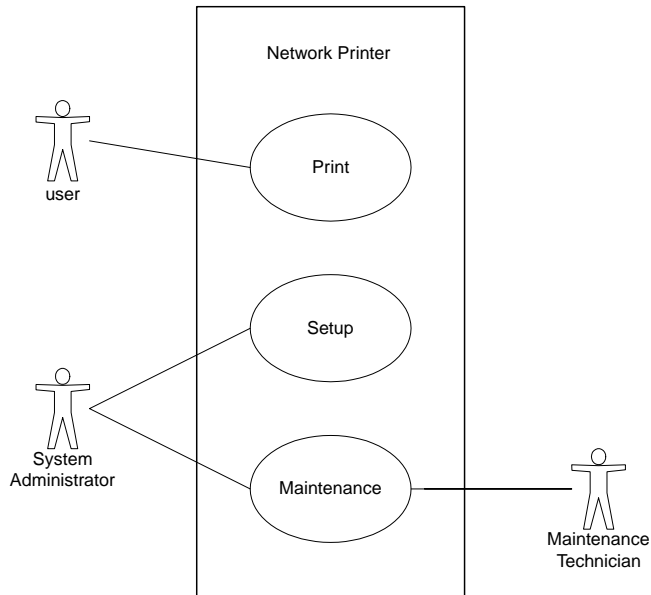



FIGURE 22.2 Use case diagram for a network printer.

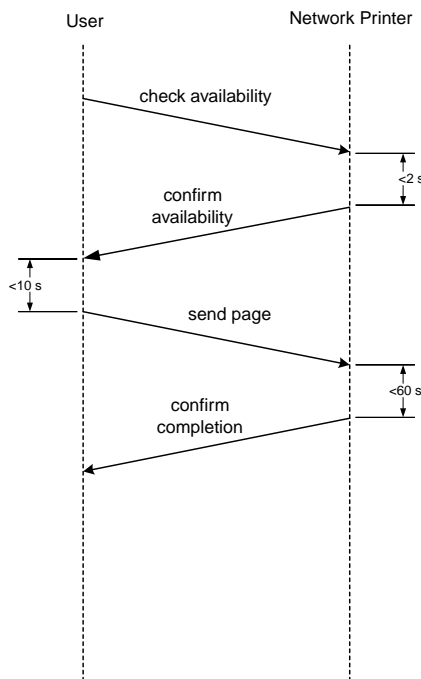



FIGURE 22.3 A sequence diagram for network printer use.

provide information about the nature of the transaction, but rather documents its existence. It serves as a top level model of a system only.

Sequence diagrams can be drawn to describe the details of individual use cases in order to clarify the interactions that must occur, timing constraints, and interactions between the system and multiple elements of the environment. Figure 22.3 shows an example of a sequence diagram for the "print" use case of Fig. 22.2.

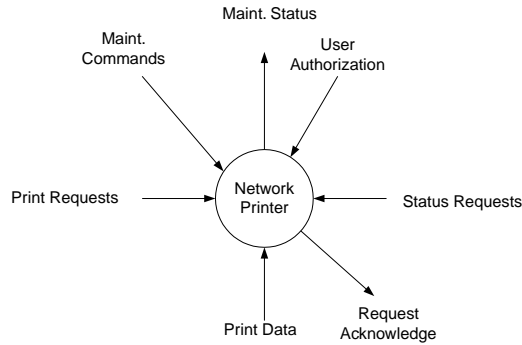



FIGURE 22.4 Context diagram for network printer.

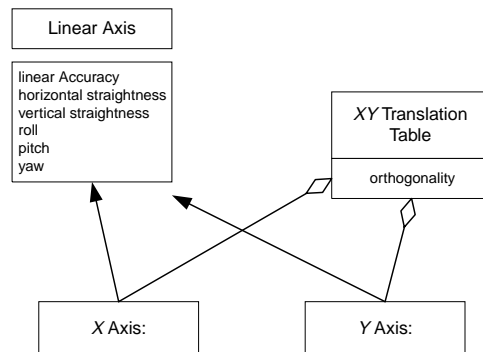



FIGURE 22.5 Class diagram for linear translation stages.

Note that the sequence diagram documents a particular instance of a particular use case. The sequence diagram can show both the direction of information flow and timing requirements. In this regard the sequence diagram and the traditional timing diagram serve similar purposes.

Context diagrams do not capture timing elements but capture the types of information that flow between the system and its environment.[5] While use case diagrams focus attention on the scenarios of system use, context diagrams focus attention on information flows that must exist to enable those scenarios. Both serve useful and complementary purposes. Figure 22.4 shows a context diagram for an inspection system like the example mentioned above. Notice that the context diagram summarizes information flow shown in the various interaction diagrams for all use cases. Regulatory and safety requirements are often in the form of limits on the interactions the system may have with aspects of its environment. These requirements must all be captured for use in later design phases and communicated back to the source of the requirement for verification of accuracy. The context diagram is also used as the top level of a data flow diagram for a system if structured analysis is used.[6]

Just as the context diagram ties to structured analysis techniques, object oriented analysis techniques provides the class relationship diagram as a means of documenting the relationships between a system and its environment and between components of a system.[7] Class relationship diagrams show how a system is composed of subsystems, how components are similar to one another, and how they differ. For example, Fig. 22.5 shows a class relationship diagram for a two-axis translation system consisting of two single axis subsystems. This diagram documents the existence of 13 error components that must be specified in the requirements phase. It does this by showing that the *X* and *Y* axis components are instances of a class of single axis translation stages, that each have six error components, and that they are components of a

system that adds a single additional error. By documenting relationships such as composition and inheritance, requirements and the interdependence of requirements can be represented in a compact and comprehensible way.

One key aspect of the requirements definition phase is the importance of defining requirements without specifying a preferred solution embodiment. Hence, modeling methods should be chosen to document these requirements and their intrinsic relationships without implying a particular solution.

Phase 2

In the concept generation phase, our objective is to generate multiple design concepts that might satisfy the requirements identified in phase 1. Here we need modeling techniques that allow us to describe possible solutions with varying levels of detail dependent on the degree of detail needed to document the key elements of the concept. Since individual concepts generated in this phase may only satisfy some portions of the design requirements, it is critical that modeling at this point allow for partial descriptions of embodiments and for the easy combination of design concepts. For this reason, our models must clearly document the portions of the requirements satisfied as well as any unspecified parameters or additional requirements introduced in the concept. For some problems, block diagrams showing interconnections between components solving portions of the problem are a useful modeling tool at this stage. Figure 22.6 shows two possible block diagrams describing a given design concept. The first specifies a particular control algorithm, sensor, and actuator, while the second leaves these particulars unspecified and simply describes a closed loop controller. Depending on the situation either of these may be appropriate descriptions. The first provides details and is closer to a complete design while the second, being more generic, is easier to combine with other concepts to generate hybrid solutions.

Block diagrams are not the only modeling tool appropriate at this stage. For other problems, schematics showing arrangements of components or equations or psuedocode of proposed algorithms may be employed. As this phase continues, design concepts are often combined to form potential solutions to the overall design problem; therefore, it is useful if the model of each concept can contain descriptions of preconditions for its use, results, and other parameters that help a design team determine how to combine concepts. Similarly, once potential solutions are formed by concatenating concepts, it is often useful to clean up the final concept by combining features from different component concepts or eliminating overlapping features that are no longer needed.[8] This process is facilitated by modeling techniques that allow simultaneous modeling of mechanical, thermal, electrical, and software components and concepts. These techniques include the familiar linear graph and bond graph models.[9,10]

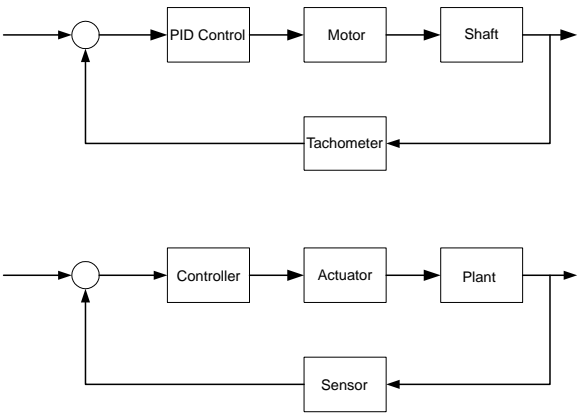



FIGURE 22.6 Block diagrams at two levels of detail.

Phase 3

In the third phase we evaluate potential solutions in terms of the problem requirements. This phase is eased if we have a model that allows us to compare problem requirements with design features. A particular design methodology may include various quality criteria in addition to those set by customer requirements and the regulatory environment, e.g., the design axioms of Suh prefer designs with minimum information content and which satisfy each requirement by independent features of the design.[11]

One approach to evaluation is to attempt to find numerical criteria for all requirements and evaluate solutions via minimization of an overall cost function (or equivalently, the maximization of an overall value function). This approach has proven effective in certain types of problems where requirements are amenable to quantification (for example, the optimal solution being the one that meets requirements with minimal weight or economic cost). Even in these cases, it can be difficult to determine the relative importance of diverse requirements in the formulation of the value function. For example, it may be desirable for a design to have minimal parts cost and minimal weight. An appropriate cost function might be

$$\Psi = \sqrt[n]{\left(\frac{c}{c_0}\right)^p + \left(\frac{w}{w_0}\right)^q}$$

where c and w are the cost and weight, respectively, and c_0 and w_0 are scaling factors indicating when the two factors are of equal importance, while the exponents p, q, and n express the relative importance of minimizing the two factors. Clearly, there are many possible choices for these parameters and many other models that can be used. This highlights the difficulties of this method. An example of this difficulty in a relatively simple case is determining the cost of a component being out of tolerance. Models for this problem range from a step function of cost = 0 within tolerance and cost = C outside tolerance to Taguchi's quadratic cost function.[12] Needless to say, the "optimal" solution found is typically dependent on the cost function chosen. On the other hand, in some cases, the optimum found is not strongly dependent on precise choice of cost function and a satisfactory evaluation can be made without expending excessive effort to obtain an exact cost function. In fact, if there are relatively few design options from which to choose, it is possible to invert this problem. Rather than finding the design that minimizes a particular cost function, one can instead find the range of cost function parameters that result in any particular solution being optimal. It is often easy to determine which range of parameters is most realistic.

Phase 4

The fourth design phase is detailed design, in which the entire process is repeated to resolve open design details for the individual components of the resulting design. This is in keeping with the design heuristic "allocate resources as long as the cost of not knowing exceeds the cost of finding out."[13] The process is inherently recursive, with each high level design decision producing a simpler design problem at a lower level of abstraction with simpler requirements. To accommodate this, we need to be able to model the design at multiple levels of abstraction and to allow the specification of interfaces between components. These models must allow us to define static as well as dynamic (behavioral) components of the interface.

In addition to this recursive aspect of design, there is of course iteration also possible, as dead ends are discovered in an otherwise promising chain of design decisions and as changes in the marketplace demand revisions of a product. To accommodate this iteration, design models should document not only decisions made, but also reasons for those decisions. Otherwise, the reasoning must be rediscovered in each iteration, or worse, may not be rediscovered until it is too late.

22.2 The Goals of Modeling

We have seen that models serve many purposes in mechatronics design. First, models serve to document concepts, assumptions, and requirements and allow the communication of these concepts to others in the design group and the various stakeholders of the design process. Second, models provide a hierarchical framework which allows division of labor in the design process and permits concurrent work on separate components.

Third, models can provide us with insights into the behavior of a system which might be obscured when we attempt to see the system in its full complexity. Fourth, they allow us to grasp similarities with other systems and use prior experience to help solve current problems. Fifth, models provide us a way of identifying and testing our ignorance of the actual behavior of the system by identifying unknown parameters and providing hypotheses worthy of testing. In this section, we shall consider these diverse purposes of modeling and examine the characteristics that cause a particular modeling methodology to be better or more poorly suited for each of these purposes.

Documentation and Communication

The first purpose of a model that we shall consider is documentation. We often forget that aside from the various purposes of modeling in the design process, our models are often our best tools for communicating with our colleagues, our customers, and our successors. Every document generated in the design process is a model of some part of the system. To the extent that we can avoid double documentation of the same concept, we not only reduce workload and decrease time to market, but we also reduce the likelihood of inconsistencies in our documentation. How do our models serve as documentation? Consider a circuit diagram. It is not the circuit, but documentation of the circuit's design. It is also a model of the behavior we are seeking from the circuit. Nonidealities and component variations will cause the actual circuit to behave differently from this model, but the model still serves its purpose. Earlier in the design process, requirements documents also document the product, but in terms of requirements rather than the features that satisfy them.

Design documentation is important for at least four distinct groups. The engineering team itself needs communication among its members to ensure that effort is not wasted on elements that do not interoperate or that do not contribute to the overall product meeting requirements. Communication is required between the design team and the customer to ensure that requirements are correctly understood and implemented in the design. Management needs to understand the status of the project in terms of outstanding risks, ongoing expenses, and tradeoffs that would affect market size. Finally, future design teams can benefit from documentation of earlier design concepts, analyses, and decisions. This role of documentation as a communication with future workers in the field has been recognized as a key part of all scientific disciplines[14] but is often overlooked by engineers focused on meeting short-term deadlines.

What characteristics of a model lead to good documentation and clear communication? Clearly, the model must be understood by others. This requirement encourages the use of standardized modeling techniques whenever possible. It also encourages the use of modeling techniques that are interdisciplinary whenever feasible. When choosing models for documentation purposes, one must remember that each model highlights certain features of the design while hiding others. For example, the sequence diagram clearly documents communication between two objects while hiding details of how the objects generate outgoing messages or interpret incoming ones. If multiple models of the system are used to document the various aspects of the system, there should be some way to cross-reference items between models (e.g., determine that timing requirements documented in a sequence diagram model are met by the dynamics of a system described in a block diagram). Also, if the documentation is to be useful to future designers, it should be possible to cross reference design models based on a variety of criteria including system requirements and included components, so that future designs can benefit from current ones. Clearly, any modeling technique that serves some other purpose in design will also serve as documentation, but the model must be chosen so as to match the features one wishes to document.

Hierarchical Framework

Real mechatronic systems include a large number of components that interact in many ways. Our models simplify these complex interfaces and describe products and processes in terms of simply interacting components with well-defined interfaces and behaviors. This simplification allows us to subdivide complex problems into a set of simpler problems whose solutions can be easily integrated. In complex mechatronic systems where we must consider components that interact in several energy domains, this

simplification is essential. When we draw a block diagram of a system we are dividing the system's behavior into a set of defined elements with modeled behavior and a known set of interactions between those elements, which is itself amenable to analysis. This role of modeling would lead us to select modeling techniques that allow us to divide the model into portions that can be independently tackled by independent engineers. For example, systems of linear differential equations may give us great insight into the fundamental modes of a system, but give us little insight on how to divide a design problem among members of a team. However, a block diagram, a class diagram, or a data flow diagram provides clear interfaces between elements that allow for subdivision.

Insights

Regarding insights into the behavior of the system, different models provide different types of insights. Signal flow graphs of electromechanical systems show the presence of feedback loops that stabilize or destabilize the system. Differential equations provide insights regarding the time scales of various behaviors and the relative importance of various factors as well as providing estimates of the validity of simplifying assumptions. Similarly, sequence and timing diagrams can provide insight regarding the processing power required to meet timing requirements for various use cases. In each case the model provides insights into the problem or the characteristics of a proposed solution by bringing into focus certain aspects of the modeled system while hiding others from view.

Analogies

Related to the insights a model can give us regarding the system under consideration are the insights that a model can provide in allowing us to see similarities to other systems that we or others have seen before. In this regard, modeling techniques that use analogies between various domains can be useful. However, analogies must be chosen with care. Consider the common analogy between electrical and mechanical quantities shown in the left half of Fig. 22.7. In this analogy, velocity is analogous to voltage and force is analogous to current. However, as seen in the right half of Fig. 22.7, the equations for the electrical system are unchanged in the dual system, in which current and voltage are exchanged and the roles of inductance and resistance are exchanged with capacitance and conductance, respectively. This dual system results in a different mechanical analogy, in which velocity is analogous to current and force is analogous to voltage. Each of these analogies can prove useful in moving design parameters from one energy domain to another, as the duality between the two electrical models can prove useful in circuit design. With any of these analogies, it must be remembered that real components deviate substantially from their idealized models and that the analogies do not strictly hold. This can be both a bane and a blessing, since a design that suffers from component nonidealities might be replaced by an analogous design from a domain with different but less detrimental nonidealities.

Element	Electrical	Mechanical	Electrical	Mechanical
Capacitance	$i = C\frac{dv}{dt}$	$f = M\frac{dv}{dt}$	$v = \frac{1}{C}\int idt$	$f = k\int vdt$
Inductance	$i = \frac{1}{L}\int vdt$	$f = k\int vdt$	$v = L\frac{di}{dt}$	$f = M\frac{dv}{dt}$
Resistance	$i = \frac{1}{R}v$		$v = Ri$	$f = Bv$
Conductance	$i = Gv$	$f = Bv$	$v = \frac{1}{G}i$	

FIGURE 22.7 Electrical–mechanical analogies.

Identification of Ignorance

The final role of modeling is to help us identify our ignorance. This takes two forms. First, the construction of the model often requires us to name parameters whose numerical value is unknown. This leads us to estimate the possible range of such parameters or to design experiments to determine the parameter's value. Identifying missing details is the first step in determining them. In this way models help us identify our ignorance of details of the problem or our proposed solution. Secondly, the predictions we make with the help of our models are testable. The adequacy of a model to describe system behavior can be determined by comparing actual system behavior with predictions based on that model. Differences between actual system performance and a model's predictions point to errors in the model that may be significant. By testing the validity of our models we identify aspects of the real system that are inadequately reproduced in our model or artifacts of the model that do not exist in the real system. These discrepancies may point to an inappropriate assumption or to a fundamental misunderstanding of the physical system. In either case, the source of these discrepancies should be understood in order to determine if our model is adequate for its intended purpose.

22.3 Modeling of Systems and Signals

In the area of systems and signals we typically deal with the modeling of system dynamics and information flows through the system and its environment. As can be seen from the discussion above, this is but a small part of the modeling needed for system design. However, there are a number of decisions that must be made in selecting a model for the behavior of a dynamic system. These include the choice between analytic and numerical modeling, the use of distributed or lumped parameters, the approach used to model random factors, and the choice between linear and nonlinear models. These issues are considered in the paragraphs below.

Analytical vs. Numerical Models

In this area both analytical and numerical tools are available to support our efforts, so we have a number of decisions to make regarding the level of complexity of the models we choose to use for various tasks. Analytical tools provide insights into overall system behaviors and can allow us to make comparisons between dissimilar systems based on a similarity in their models. These analogous systems often prove useful in the concept generation phase of design because they allow us to bring to bear solutions to diverse problems to the solution of the problem at hand. However, as described above, these analogies usually break down upon closer inspection. In the case of analogies in dynamics, they are typically based on systems described by equivalent differential equations (e.g., electronic and mechanical oscillators). However, these linear differential equations are but simplifications that ignore nonlinearities and time-dependent behavior that may greatly affect system behavior. Therefore, when applying analogies one must always be careful to explicitly state assumptions made in the modeling process.

Numerical tools, on the other hand, do not yield the same insights of analytical tools, but provide other and more detailed insights into real system behavior and performance. Typically these tools become most useful later in the design process, both for tradeoff analysis and for detailed design work. The availability of these tools opens many doors to modeling options that have in the past been closed and allows us to consider options that have historically been considered unwieldy. For example, numerical solution of nonlinear differential equations allows modeling of systems that had previously been approached only with linear approximations.

Partial vs. Ordinary Differential Equations

One modeling choice that must be made is whether we must model the dynamics of the system using ordinary or partial differential equations. Partial differential equations must be used whenever values of interest vary spatially as well as temporally. For example, if we are considering the design of a robot arm, we are free to use ordinary differential equations if we consider the arm to be rigid, partial differential

equations if we wish to describe the bending of a flexible arm in detail. However, what if we recognize a flexibility in the arm but do not need to know the details of its motion but only the effect of its flexure on the end effector? In this case we can make a lumped parameter model of the arm that summarizes its dynamics in terms of end effector motion and dynamic forces on the driver and end effector. When can this lumped parameter model be used? There are two restrictions. First, the details of the variation of behavior over space must be uninteresting to us. If we need to know these details (for example, to determine stress concentration, or temperature distribution) we cannot apply the simplified model. Secondly, the partial differential equation may not be amenable to forming a lumped form due to its complexity.

Stochastic vs. Deterministic Models

One choice that must be made in the modeling of system behavior is how uncontrolled variability will be represented in the model.[15] The approach taken depends a great deal on the source and characteristics of the variability. For example, the values of some system parameters may not be precisely known, but may be constant over time. Others may vary slowly over time in unknown ways. In the first case, the sensitivity of the system to parameter value can be determined, using a variety of both analytical and numerical techniques. In the latter case, one must examine the speed of parameter variation to determine if the system can be analyzed in quasi-steady state or if the dynamics of the parameter variations are coupled with the dynamics of the system. For this purpose, it is often useful to write equations in dimensionless form where time and scale constants indicate critical speeds for coupling.

When the variation of parameters over time must be considered or when some inputs to a system have uncontrolled variability, we must choose whether to model the input as an arbitrary signal or a signal with known statistical parameters. In the first case, we have just another input and can analyze the system for sensitivity to this input. In the second case, we can characterize the system in terms of autocorrelation or equivalent spectral measures and use optimal filtering techniques to reduce uncertainty.[16] A key point in the statistical modeling of a system is to base the model on the known variability of the process rather than assuming additive, white, Gaussian noise at every turn.[17]

Linear vs. Nonlinear

A wealth of techniques are available for the analysis of linear time invariant systems. Unfortunately, the world gives us nonlinear components. One advantage of modeling techniques that allow us to divide a complex system into simpler sub-components is that we can often isolate nonlinearities as individual elements and develop linear models for them. For small variations about a differentiable point on a nonlinear curve, we can always find a linear model, but the value of this model might be extremely limited depending on its accuracy over a reasonable range of input values. In the case of a discontinuous or nondifferentiable nonlinearity, there is a greater problem. A linear model can be found only when we are far from the discontinuity or when the discontinuity is small compared to the linear portion of the model.

However, nonlinearities also add essentially new behaviors to systems that are not possible in purely linear systems. For this reason alone, nonlinear models are needed in certain circumstances. As an example, for a linear system, stable oscillations are only theoretically possible for differential equations with purely imaginary eigenvalues. Positive real parts lead to growing oscillations, while negative real parts lead to damped oscillations. Clearly, the behavior of such a system is highly sensitive to system parameters. Small variations will either kill the oscillation or drive the system into nonlinear regions of operation. Furthermore, even for perfect oscillators, the amplitude of oscillations is unconstrained. However, a nonlinear system such as the Van der Pol equation:

$$\ddot{y} = -\omega^2 y + \alpha\left[1 - \left(\frac{y}{y_0}\right)^2\right]\dot{y}$$

includes a damping term that increases with the magnitude of the oscillation. For oscillations small compared to y_0 the oscillations grow, while larger oscillations are damped. This and other essentially

nonlinear behavior cannot be modeled adequately with linear models. Although once this behavior is recognized through the use of a nonlinear model, the resulting steady-state system (a stable oscillator) can be modeled for many purposes using its linear (although physically unrealizable) equivalent.

When using differential equations to model a system, we must be particularly careful in our choice of units of measure and in how we group parameters. For example, in the Van der Pol equation above, we can choose an alternative scale for y such that y_0 is 1 and a time scale such that ω is 1. In this case we have

$$z'' = -z + \varepsilon(1 - z^2)z'$$

where $z = y/y_0$, $\tau = \omega t$, and $\varepsilon = \alpha/\omega$. This scaling provides a dimensionless equation as well as providing an indication of the system's natural time scale, the scale of y, and an indication that the size of α relative to ω is important. In general, we find that writing equations in dimensionless form provides three advantages. First, as mentioned above, the scaling factors provide a sense of time scale and magnitude of various features of the system's dynamic behavior. Second, the dimensionless parameters that result from combining physical system parameters provide a way of characterizing the behavior of a wide range of systems in terms of parameters that are easily mapped. Third, by putting the equation into dimensionless form, we are able to categorize it and recognize similar equations in different domains.

References

1. Hubka, V. and Eder, W.E., *Engineering Design—General Procedural Model of Engineering Design,* Heuista, Zurich, 1992.
2. Pahl, G. and Beitz, W., *Engineering Design: A Systematic Approach,* Springer-Verlag, Berlin, 1988.
3. Dym C.L., *Engineering Design—A Synthesis of Views,* Cambridge University Press, New York, 1994.
4. Booch, G., Jacobson, I., Rumbaugh, J., and Rumbaugh, J., *The Unified Modeling Language User Guide,* Addison-Wesley, New York, 1998.
5. Douglass, B., *Doing Hard Time: Developing Real-Time Systems with UML, Objects, Frameworks and Patterns,* Addison-Wesley, New York, 1999.
6. Demarco, T., *Structured Analysis and System Specification,* Yourdon Press, NJ, 1978.
7. Coad, P. and Yourdon, E., *Object-Oriented Analysis,* Yourdon Press, NJ, 1990.
8. Glegg, G.L., *The Design of Design,* Cambridge University Press, London, 1969.
9. Shearer, Murphy, and Richardson, *Introduction to Dynamic Systems,* Addison-Wesley, Reading, MA, 1967.
10. Karnopp, D. and Rosenberg, R., *System Dynamics: A Unified Approach,* Wiley, New York, 1975.
11. Suh, N.P., *Principles of Design,* Oxford University Press, NY, 1990.
12. Taguchi, G. and Yokoyama, Y., *Taguchi Methods: Design of Experiments,* American Supplier Institute, Dearborn MI, 1994.
13. Koen, B.V., *Definition of the Engineering Method,* American Society for Engineering Education, Washington, 1985.
14. Lonergan, B., *Method in Theology,* University of Toronto, Toronto, 1990.
15. Zhou, K., *Essentials of Robust Control,* Prentice-Hall, NJ, 1998.
16. Papoulis, A., *Probability, Random Variables, and Stochastic Processes,* McGraw-Hill, New York, 1991.
17. Law, A. and Kelton, W., *Simulation, Modeling and Analysis,* McGraw-Hill, New York, 1991.

23

Signals and Systems

Momoh-Jimoh Eyiomika Salami
International Islamic University of Malaysia

Rolf Johansson
Lund Institute of Technology

Kam Leang
University of Washington

Qingze Zou
University of Washington

Santosh Devasia
University of Washington

C. Nelson Dorny
University of Pennsylvania

23.1 Continuous- and Discrete-Time Signals

Signals are physical variables or quantities measured at various parts of a system, which when processed yield the desired information. A wide variety of signals are often encountered in describing many practical systems. Electrical signal, in form of current and voltage, is the most easily measured quantity, hence the need to use sensors and transducers to transform other non-electrical quantity into electrical signals. These signals must be processed by appropriate techniques if desirable results are to be obtained. Several methods of signal representation, suitable for effective signal processing in both time and frequency domains, are discussed in this section.

Signal Classification[1–4]

Signals are broadly classified as either continuous-time (CT) or discrete-time (DT) signals, and each of these may in turn be categorized as deterministic or random signals. A deterministic signal can always be expressed mathematically, whereas the time of occurrence or value of a random signal cannot be predicted with certainty. A CT signal, $x(t)$, has a specified value for every value of time, t, while a DT signal, $x(n)$, has specified a value only at discrete points, that is, for integer values of n. Closely related

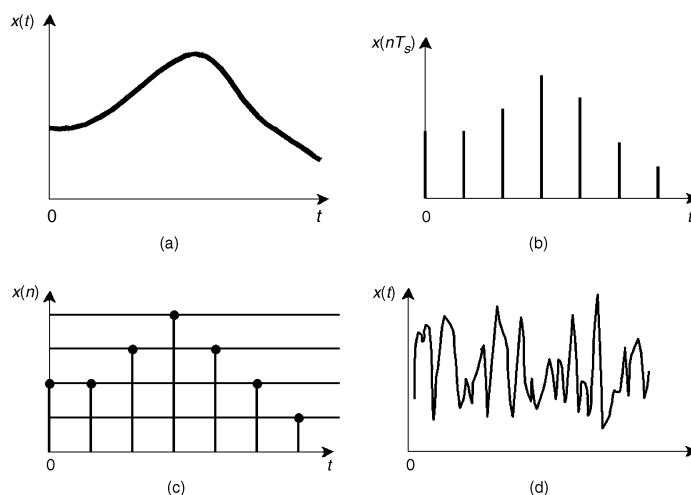



FIGURE 23.1 Description of some classes of signals: (a) continuous-time analog, (b) sampled-data, (c) digital signal, (d) random signal.

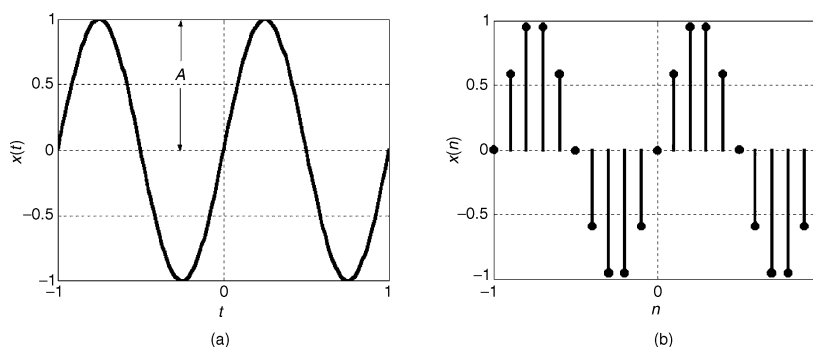



FIGURE 23.2 Description of periodic signals: (a) CT, (b) DT.

to CT and DT signals are analog and digital signals, respectively. If the amplitude of a signal can take on any value in a continuous range, then it is an analog signal. On the other hand, the amplitude of a digital signal can have only finite number of values at discrete points. Examples of continuous-time, discrete-time, digital, and random signals are shown in Fig. 23.1.

Deterministic signals fall into two main categories, namely periodic and aperiodic signals. A periodic signal has the same values at times separated by one period, T, that is, $x(t)$ satisfies the relation, $x(t) = x(t + T), -\infty < t < \infty$. An example of a CT periodic signal is a sinusoidal waveform of the form $x(t) = A \sin(\Omega t + \theta)$, where θ is the phase in radians, $\Omega = 2\pi F$ is the frequency in radians per second, and F is the frequency in hertz. It should be noted that the frequency range for the analog sinusoids is $-\infty < F < \infty$. A periodic DT signal is described by $x(n) = x(n + N), -\infty, < n < \infty$, where N represents the period. An example of this is the sinusoidal waveform $x(n) = A \sin(2\pi rn + \theta), -\infty < n < \infty$, where r is the signal frequency per sample frequency and has values in the range $-\frac{1}{2} \leq r \leq \frac{1}{2}$. Samples of sinusoids having frequencies within this range are unique and distinct. However, DT sinusoids whose frequencies are separated by an integer multiple of 2π are identical. Examples of analog and DT sinusoids are depicted in Fig. 23.2.

Any deterministic signal that is not periodic is referred to as aperiodic or nonperiodic. Damped sinusoids and exponential decaying signals are common examples of aperiodic signals. For some applications, it may

be quite useful to classify the signals according to their energy or power content. The total energy over the range $t \in (-\infty, \infty)$ for a CT signal is given by

$$E = \lim_{T \to \infty} \int_{-T/2}^{T/2} |x(t)|^2 dt \tag{23.1}$$

and the average power is defined as

$$P = \lim_{T \to \infty} \left[\frac{1}{T} \int_{-T/2}^{T/2} |x(t)|^2 dt \right] \tag{23.2}$$

Consequently, $x(t)$ is an energy signal if and only if $0 < E < \infty$, which implies that $P = 0$. Similarly, $x(t)$ is a power signal if and only if $0 < P < \infty$, indicating $E = \infty$. A signal that fails to satisfy either definition is, therefore, neither energy nor power signal. These definitions are also applicable to DT signals except that the integral in Eqs. (23.1) and (23.2) is replaced by summation. In general, periodic signals exist for all the time and as such have infinite energy. However, they have finite average power, hence they are power signals. On the other hand, bounded finite-duration signals are energy signals. The classification of a signal to finite energy, finite power, or neither is important so that appropriate and effective procedures can be selected for its analysis.

Singularity Functions

Singularity functions are useful for signal modeling, that is, they serve as basis for representing complex signals to simplify their analysis.

The Unit Impulse Function

The impulse or delta function is a mathematical model for representing physical phenomena that occurs within very small time duration; this time duration can be assumed to be equal to zero. The unit delta function is not a mathematical function in the usual sense; rather it is a distribution or a generalized function. Thus, the impulse function can be described by its effect on the *test function* $\phi(t)$, that is,

$$\int_{-\infty}^{+\infty} \phi(t)\delta(t)dt = \phi(0) \tag{23.3}$$

provided $\phi(t)$ is continuous at $t = 0$. This equation shows the shifting property of the delta function. A graphical plot of the delta function is shown in Fig. 23.3(a). Table 23.1 shows the operating properties of the delta function.

The Unit Step Function

The unit step function is particularly useful for the mathematical analysis of CT signals. This is depicted in Fig. 23.3(b), and is defined as

$$u(t) = \begin{cases} 1, & t > 0 \\ 0, & t < 0 \end{cases} \tag{23.4}$$

TABLE 23.1 Properties of the Delta Function

Property	Mathematical Expression
Sampling	$\int_{-\infty}^{+\infty} x(t)\delta(t-a)dt = x(a)$
Shifting	$x(t)\delta(t-a) = x(a)\delta(t-a)$
Scaling	$\delta(at \pm b) \equiv \frac{1}{\lvert a \rvert}\delta\left(t \pm \frac{b}{a}\right)$
Convolution	$x(t)*\delta(t-a) = \int_{-\infty}^{+\infty} x(\tau)\delta(t-\tau-a)d\tau = x(t-a)$

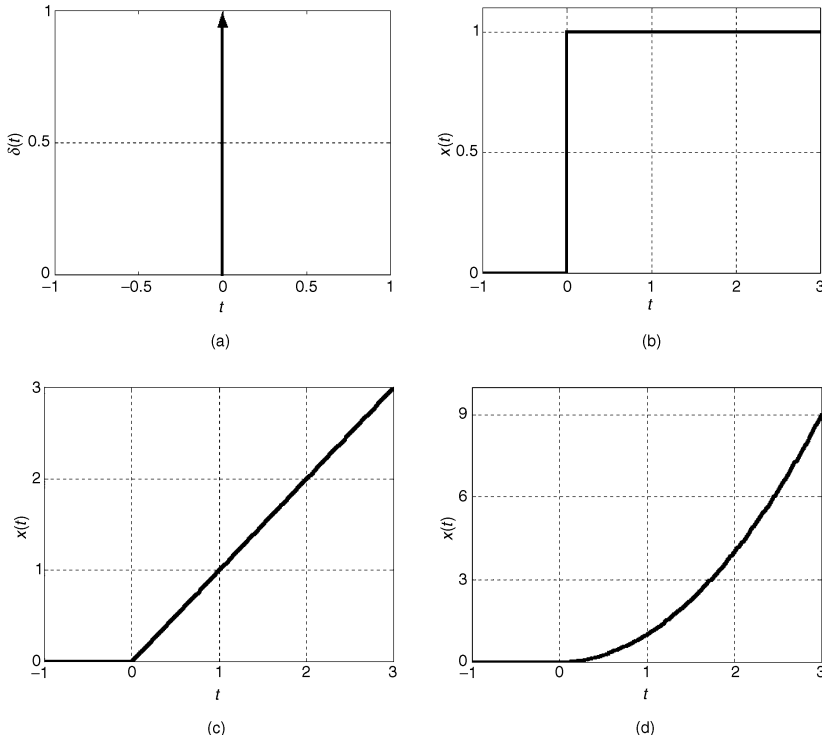



FIGURE 23.3 Description of singularity function: (a) impulse, (b) step, (c) ramp, (d) parabolic.

The signum function is derived from the step function according to

$$\mathrm{sgn}(t) = 2u(t) - 1$$

The Ramp Function

Integrating (23.4) yields the ramp function shown in Fig. 23.3(c). The unit ramp function $r(t)$ is expressed as

$$r(t) = \begin{cases} t, & t \geq 0 \\ 0, & t < 0 \end{cases} \tag{23.5}$$

Furthermore, integrating $r(t)$ yields a unit parabolic signal of the form

$$a(t) = \begin{cases} \dfrac{t^2}{2}, & t \geq 0 \\ 0, & t < 0 \end{cases} \tag{23.6}$$

This is depicted in Fig. 23.3(d).

Basic Continuous-Time Signals

Figure 23.4 shows some of the elementary signals that are often encountered in signal analysis. Some of these signals can be derived directly from the singular functions discussed above. For example, the unit rectangular pulse signal that extends from $-\tau/2$ to $\tau/2$ can be expressed as

$$\prod(t) = u\left(t + \frac{\tau}{2}\right) - u\left(t - \frac{\tau}{2}\right) \tag{23.7}$$

and this is depicted in Fig. 23.4(a).

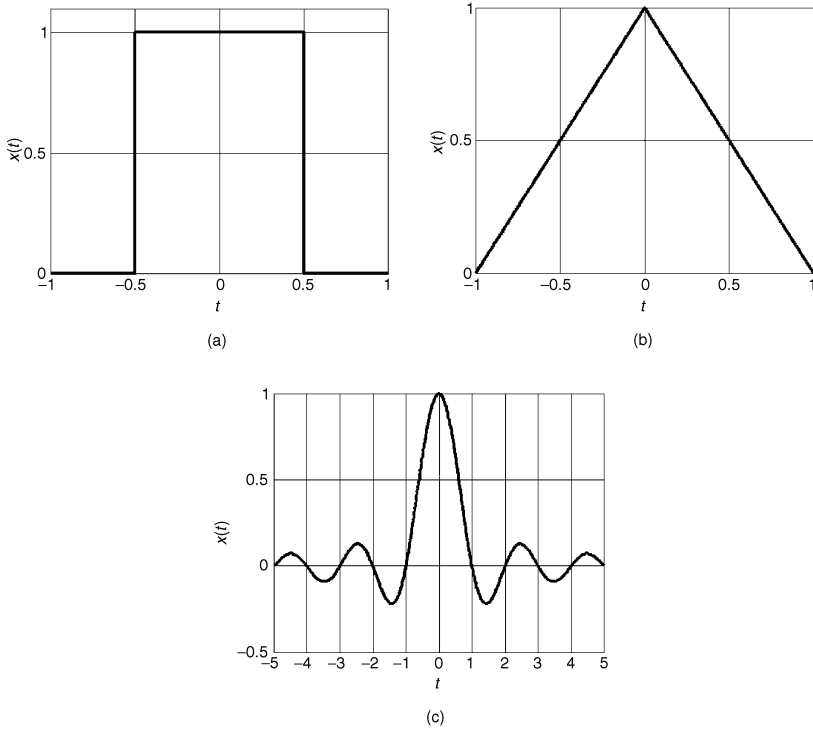



FIGURE 23.4 Description of basic CT signal: (a) rectangular pulse, (b) triangular pulse, (c) sin *c* function.

The triangular function, denoted as $\Lambda(t)$, is defined as

$$\Lambda(t) = \begin{cases} 1 - |t|, & |t| < 1 \\ 0, & \text{otherwise} \end{cases} \tag{23.8}$$

Using the convolution theorem, to be discussed in section "Analysis of Continuous-Time Signals", it can be shown that

$$\Lambda(t) = \prod(t) * \prod(t)$$

The sin*c* signal is defined as

$$\text{sin}c(t) = \begin{cases} \dfrac{\sin(\pi t)}{\pi t}, & t \neq 0 \\ 1, & t = 0 \end{cases} \tag{23.9}$$

Both the triangular and sin*c* signals are shown in Figs. 23.4(b) and 23.4(c), respectively.

Basic Discrete-Time Signals

The unit sample sequence, denoted as $\delta(n)$, is defined as

$$\delta(n) = \begin{cases} 1, & n = 0 \\ 0, & \text{otherwise} \end{cases}$$

It is also referred to as the unit impulse sequence or Kronecker delta function. The working properties of the unit sample sequence are analogous to that of $\delta(t)$ and these are shown here:

$$\sum_{n=-\infty}^{\infty} x(n)\delta(n-m) = x(m)$$

$$x(n)\delta(n-m) = x(m)\delta(n-m)$$

$$\delta(an \pm b) = \delta\left(n \pm \frac{b}{a}\right)$$

$$x(n) * \delta(n-m) = \sum_{n=-\infty}^{\infty} x(r)\delta(n-r-m) = x(n-m)$$

Note that the scaling property is only applicable when both a and b/a are integers. Two other basic signals that are useful for analysis are the unit step and unit ramp signals. The unit step sequence, $u(n)$, is defined as

$$u(n) = \begin{cases} 1, & n \geq 0 \\ 0, & n < 0 \end{cases} \tag{23.10}$$

whereas the unit ramp signal, denoted as $r(n)$, is given by

$$r(n) = \begin{cases} n, & n \geq 0 \\ 0, & n < 0 \end{cases}$$

The above three sequences are related as follows:

$$\delta(n-k) = u(n-k) - u(n-k-1)$$

$$u(n) = \sum_{m=-\infty}^{n} \delta(n-m)$$

$$r(n) = u(n) * u(n-1)$$

Figure 23.5 illustrates the above DT sequences.

Analysis of Continuous-Time Signals

Basic Operations on Signals

There are some important operations that are often performed on signals so as to understand either their characteristics or the physical phenomena generating them. The three most common operations are shifting, time scaling, and reflection. Examples of these operations are illustrated in Fig. 23.6, where $x(t)$ is expressed as

$$x(t) = \begin{cases} t+1, & -1 \leq t \leq 3 \\ 3, & 3 < t \leq 6 \\ 0, & \text{otherwise} \end{cases}$$

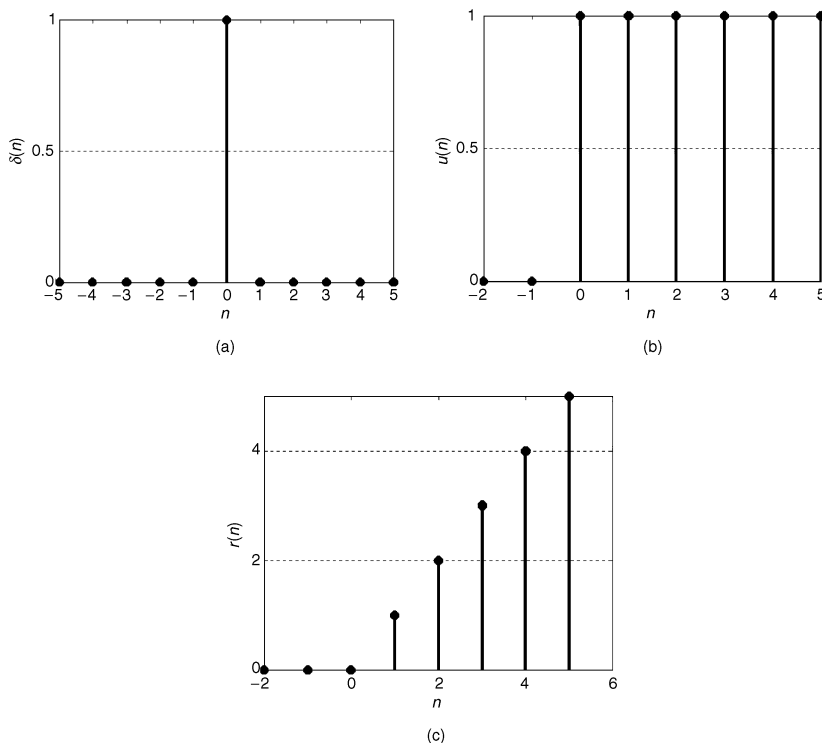



FIGURE 23.5 Description of basic DT signal: (a) unit pulse sequence, (b) unit step sequence, (c) ramp sequence.

The Convolution and Correlation Integrals[2]

Though the convolution operation is often associated with systems studies, occasionally this operation may be needed in analyzing signals obtained from a physical system. The convolution of two CT signals $x(t)$ and $y(t)$ yields $z(t)$, where

$$z(t) = x(t) * y(t) = \int_{-\infty}^{\infty} x(\tau)y(t-\tau)d\tau \tag{23.11}$$

Convolution is not limited to time-domain since it is also used to determine the frequency-domain spectrum associated with the product of two time-domain signals. The cross-correlation function between $x(t)$ and $y(t)$, denoted as $R_{xy}(t)$, is defined as

$$R_{xy}(t) = x(t) \oplus y(t) = \int_{-\infty}^{\infty} x(\tau)y^*(\tau-t)d\tau \tag{23.12}$$

Unlike the convolution there is no reflection operation here. Furthermore, the variable lag, t, is the scanning parameter that measures the degree of similarity between these two signals. If $x(t) = y(t)$, then (23.12) describes the autocorrelation function. Some properties of the correlation functions are given in Table 23.2.

Both convolution and correlation integrals are applicable to the energy as well as power signals. In the case of power signals, the integral is taken over the period T and the result is scaled by $1/T$. Correlation analysis is important as it leads to the computation of the energy spectral density for the transient signals, and power spectral density for both periodic and random signals.

TABLE 23.2 Properties of the Correlation Function

Property	Autocorrelation	Crosscorrelation
Even/Reorder	$R_{xx}(t) = R_{xx}(-t)$	$R_{xy}(t) = R_{yx}(-t)$
Upper Bound	$R_{xx}(0) \geq R_{xx}(t)$, for any t	$\lvert R_{xy}(t)\rvert \leq \sqrt{R_{xx}(0)R_{yy}(0)}$

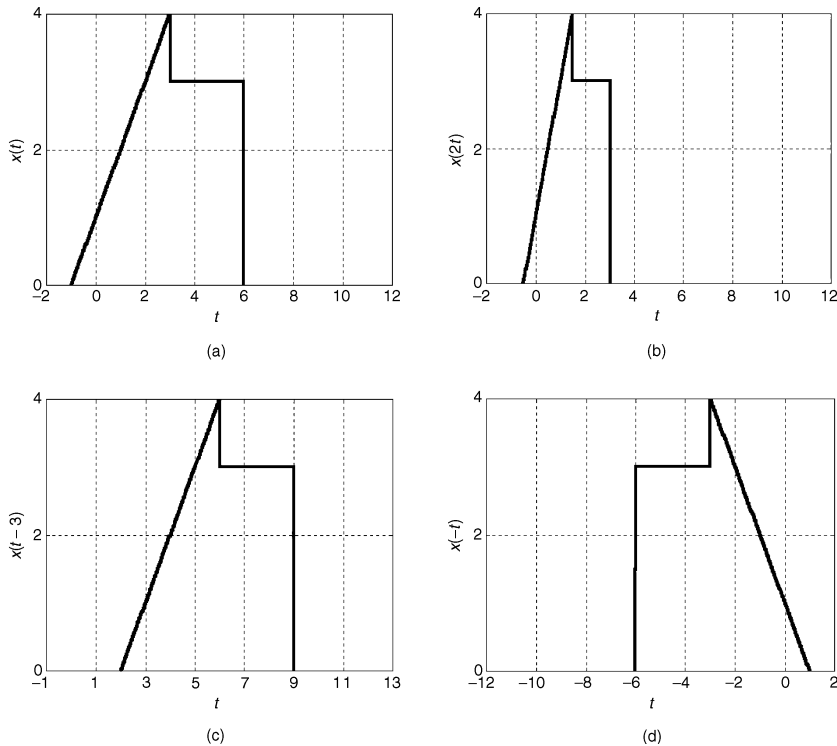



FIGURE 23.6 Basic operations on signal: (a) original signal, (b) scaling, (c) shifting, (d) reflection.

Fourier Analysis of CT Signals

So far we have discussed only the time-domain methods of analyzing CT signals. The convolution integral is of particular interest since this can be used to study how a signal is modified as it passes through a system. There is need to consider the frequency-domain methods of analysis since the convolution analysis can be laborious. Furthermore, the formulation of convolution integral is based on representing the signals by shifted δ-functions. In many applications, it is more appropriate and desirable to choose a set of orthogonal functions as the basic signals since this approach leads to a reduction in computational complexity as well as providing a graphical representation of the frequency components in a given signal.

Orthogonal Basis Functions[2,3]

It is mathematically convenient to represent arbitrary signals as a weighted sum of orthogonal waveforms as this leads to a very much simplified signal analysis as well as showing the fundamental similarity between signals and vectors. Consider a set of basis function $\phi(t)$, $i = 0, \pm1, \pm2, \ldots$. This is said to be orthogonal over an interval (t_1, t_2) if

$$\int_{t_1}^{t_2} \phi_m(t)\phi_k^*(t)dt = E_k\delta(m-k) \tag{23.13}$$

where $\phi_k^*(t)$ stands for the complex conjugate of the signal. If E_k is equal to unity for all values of k, then the $\phi(t)$ is an orthonormal set. It is relatively easy to approximate a given signal by an appropriate set of orthonormal functions as this leads to minimum error between the actual signal and its approximation. Thus, a given signal $x(t)$ with finite energy over the interval $t_1 < t < t_2$ can be expressed as

$$x(t) = \sum_{k=-\infty}^{\infty} c_k \phi_k(t) \tag{23.14}$$

where

$$c_k = \int_{t_1}^{t_2} x(t)\phi_k^*(t)dt, \quad k = 0, \pm 1, \pm 2, \ldots$$

This equation is referred to as the generalized Fourier series of $x(t)$, and the constants c_k, $k = 0, \pm 1, \pm 2, \ldots$, are called the Fourier series coefficients with respect to the orthogonal set $\{\phi(t)\}$.

Denoting the first M terms in Eq. (23.14) as $\hat{x}(t)$, the resulting error function is

$$e_M(t) = x(t) - \sum_{k=0}^{M} c_k \phi_k^*(t) \tag{23.15}$$

By computing the average power of this error function and setting its derivatives with respect to c_k to zero, yields an optimal set of $\{c_k\}$ that minimizes the error energy. Also, if $\lim_{M\to\infty} \hat{x}(t) = x(t)$, then the basis functions are said to be complete, that is, the error energy is equal to zero. When dealing with periodic signals, the time interval (t_1, t_2) is equal to the period, T, of the signal. In addition, $\phi_n(t) = \exp\{jn\omega_0\}$ is often selected as the set of basis functions, for $n = 0, \pm 1, \pm 2, \ldots$, and $\omega_0 = 2\pi/T$. The methods of computing the Fourier series coefficients are subsequently discussed.

The Complex Exponential Fourier Series

Let the signal $x(t)$ be such that $x(t) = x(t + T)$ and that it satisfies the Dirichlet conditions:[3]

1. $x(t)$ is absolutely integrable over its period, that is,

$$\int_{t_1}^{t_1+T} |x(t)|\,dt < \infty$$

2. the number of maxima and minima of $x(t)$ in each period is finite,
3. the number of discontinuities of $x(t)$ in each period is finite,
   then $x(t)$ can be expanded as

$$x(t) = \sum_{k=-\infty}^{\infty} c_k e^{jk\omega_0 t}, \quad \omega_0 = 2\pi f_0 \tag{23.16}$$

where

$$c_k = \frac{1}{T}\int_{t_1}^{t_1+T} x(t)e^{-jk\omega_0 t}dt \tag{23.17}$$

for any arbitrary value of t_1.

The coefficients c_k are called the complex Fourier series coefficients for the signal $x(t)$, which, in general, may be complex numbers.

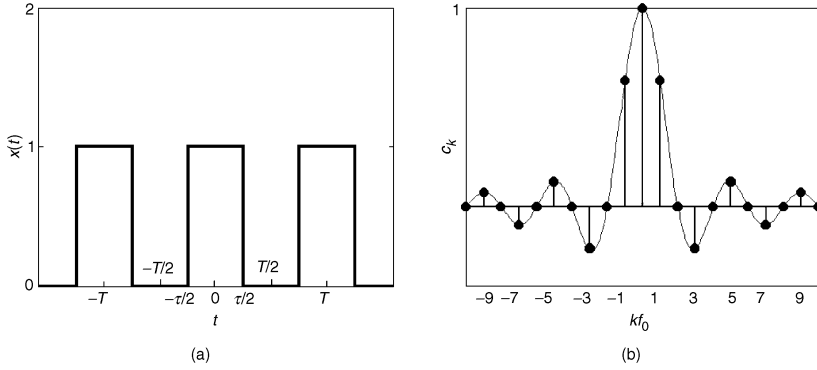



FIGURE 23.7 Rectangular periodic signal and its Fourier series coefficients.

The Dirichlet conditions are only sufficient conditions for the existence of the Fourier series expansion. The Fourier series expansion of signals, which does not satisfy these conditions, can still be obtained.

The complex Fourier series coefficients can be determined by either directly evaluating the integral in (23.17) or using the method of differentiation. The latter method relies on differentiating $x(t)$, for a certain number of times, to produce a train of impulses. These two methods of determining c_k will now be illustrated.

Consider the periodic signal shown in Fig. 23.7(a), which can be expressed as

$$x(t) = \frac{1}{T}\sum_{n=-\infty}^{\infty}\prod\left(\frac{t-nT}{\tau}\right)$$

Substituting this in Eq. (23.17) we have

$$\begin{aligned} c_k &= \frac{1}{T}\int_{-T/2}^{T/2} x(t)e^{-j2\pi k f_0 t}dt \\ &= \frac{1}{T}\int_{-\tau/2}^{\tau/2} e^{-j2\pi k f_0 t}dt \\ &= \frac{1}{\pi k}\sin(k\pi f_0\tau) \\ &= \tau f_0 \operatorname{sinc}(k f_0 \tau) \end{aligned}$$

Thus,

$$x(t) = \sum_{n=-\infty}^{\infty} \tau f_0 \operatorname{sinc}(k f_0\tau)e^{j2\pi k f_0 t}$$

A graph of the magnitude and phase spectra of the complex Fourier series coefficients are shown in Fig. 23.7(b) for varied τ/T.

Consider the computation of the complex Fourier series coefficients of the signal shown in Fig. 23.8(a) using the method of differentiation. This signal is expanded according to Eq. (23.16). Differentiating this equation twice with respect to t yields

$$x''(t) = \sum_{k=-\infty}^{\infty}(j2\pi k f_0)^2 c_k e^{j2\pi k f_0 t}$$

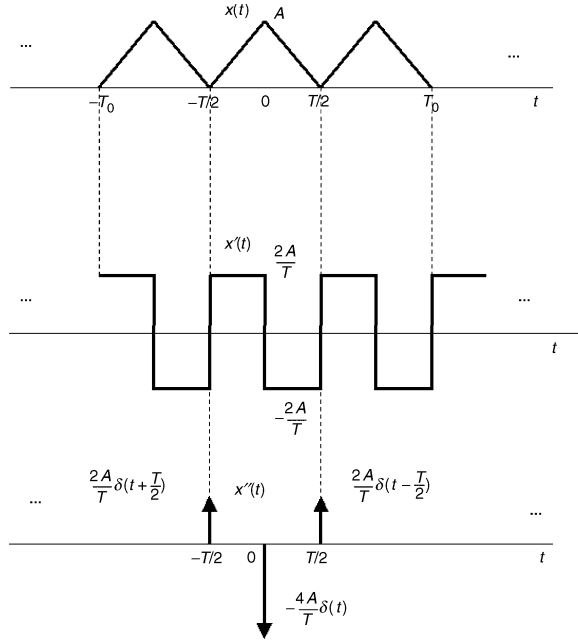



FIGURE 23.8 Illustration of the differentiation technique for Fourier series coefficient computation.

which can be written as

$$x''(t) = \sum_{k=-\infty}^{\infty} \beta_k e^{j2\pi k f_0 t}$$

Figure 23.8(c) shows the result of differentiating $x(t)$. It is noted that if a signal is periodic, its derivatives will also be periodic. This implies that β_k is the complex Fourier series coefficient of $x''(t)$ and can be computed from

$$\beta_k = \frac{1}{T}\int_{-T/2}^{T/2} x''(t)e^{-j2\pi k f_0 t}dt$$

where

$$x''(t) = 2\frac{A}{T}\left[\delta\left(t+\frac{T}{2}\right) - 2\delta(t) + \delta\left(t-\frac{T}{2}\right)\right]$$

Thus,

$$\beta_k = -8\frac{A}{T^2}\sin^2\left(\frac{\pi k T f_0}{2}\right) = (j2\pi f_0 k)^2 c_k$$

That is,

$$c_k = \begin{cases} \dfrac{2A}{\pi^2 k^2}, & k \text{ odd} \\ 0, & \text{otherwise} \end{cases}$$

and

$$c_0 = \frac{1}{T}\int_{-T/2}^{T/2} x(t)dt = \frac{A}{2}.$$

This method of determining c_k, where applicable, is less laborious than the direct method.

Fourier Series for Real Signals[3]

A real signal $x(t)$ that satisfies the Dirichlet conditions can be expanded by the following procedure. Recall Eq. (23.17) and replace k by $-k$, we obtain

$$c_{-k} = \frac{1}{T}\int_{t_1}^{t_1+T} x(t)e^{j2\pi kf_0 t}dt = \left[\frac{1}{T}\int_{t_1}^{t_1+T} x(t)e^{-j2\pi kf_0 t}dt\right]^* = c_k^* \qquad (23.18)$$

That is, the positive and negative coefficients are complex conjugates of each other for real signals. For such signals, $|c_k|$ has even symmetry and $\angle c_k$ has odd symmetry with respect to $k = 0$. Denote $c_k = (a_k - jb_k)/2$, then $c_{-k} = (a_k + jb_k)/2$, so that

$$c_k e^{j2\pi kf_0 t} + c_{-k}e^{-j2\pi kf_0 t} = \frac{a_k - jb_k}{2}e^{j2\pi kf_0 t} + \frac{a_k + jb_k}{2}e^{-j2\pi kf_0 t}$$

Since c_0 is real and defined as $c_0 = a_0/2$, then we can write

$$x(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty}[a_k\cos(2\pi kf_0 t) + b_k\sin(2\pi kf_0 t)] \qquad (23.19)$$

This relationship, which only holds for a real periodic signal $x(t)$, is called the trigonometric Fourier series expansion. Since

$$\begin{aligned} c_k = \frac{a_k - jb_k}{2} &= \frac{1}{T}\int_{t_1}^{t_1+T} x(t)e^{-j2\pi kf_0 t}dt \\ &= \frac{1}{T}\int_{t_1}^{t_1+T} x(t)[\cos(2\pi kf_0 t) - j\sin(2\pi kf_0 t)]dt \end{aligned} \qquad (23.20)$$

Hence

$$a_k = \frac{2}{T}\int_{t_1}^{t_1+T} x(t)\cos(2\pi kf_0 t)dt \qquad (23.21a)$$

$$b_k = \frac{2}{T}\int_{t_1}^{t_1+T} x(t)\sin(2\pi kf_0 t)dt \qquad (23.21b)$$

and

$$a_0 = \frac{2}{T}\int_{t_1}^{t_1+T} x(t)dt$$

The third method of Fourier series expansion of real signals is given by

$$x(t) = c_0 + \sum_{k=1}^{\infty} c_k\cos(2\pi kf_0 t + \theta_k) \qquad (23.22)$$

TABLE 23.3 Fourier Series Symmetry Conditions

Type of Symmetry	Real Fourier Series Coefficient	Complex Fourier Series Coefficients	Comments
Even periodic $x(t) = x(-t)$	$a_k = \frac{4}{T}\int_0^{T/2} x(t)\cos(2\pi k f_0 t)\,dt$ $b_k = 0$	$c_k = \frac{1}{2}a_k$ c_k has real value	Phase of c_k is either zero or π
Odd periodic $x(t) = -x(-t)$	$a_k = 0$ $b_k = \frac{4}{T}\int_0^{T/2} x(t)\sin(2\pi k f_0 t)\,dt$	$c_k = -j\frac{1}{2}b_k$ c_k has imaginary values	Phase of c_k is either $\pi/2$ or $-\pi/2$
Half-wave even symmetry $x(t) = x\left(t + \frac{T}{2}\right)$	a_{2k} and b_{2k} may have nonzero values but $a_{2k+1} = 0, \quad b_{2k+1} = 0$	$c_{2k} \neq 0$ $c_{2k+1} = 0$	
Half-wave odd symmetry $x(t) = -x\left(t + \frac{T}{2}\right)$	a_{2k+1} and b_{2k+1} may have nonzero values but $a_{2k} = 0, \quad b_{2k} = 0$	$c_{2k} = 0$ $c_{2k+1} \neq 0$	

TABLE 23.4 Properties of the Fourier Series

Property	Signal Description	Fourier Series Coefficients, c_k
Linearity	$ax(t) + by(t)$; a, b constants	$a\alpha_k + b\beta_k$
Multiplication	$x(t)y(t)$	$\alpha_k * \beta_k$
Convolution	$x(t) * y(t)$	$\alpha_k \beta_k$
Parseval's theorem	$\frac{1}{T}\int_{t_1}^{t_1+T} \lvert x(t)^2 \rvert\,dt$	$\sum_{k=-\infty}^{\infty} \lvert\alpha_k\rvert^2$
Time shift	$x(t \pm \tau)$	$\alpha_k e^{\pm j2\pi f_0 \tau}$
Differentiation	$\frac{d^n}{dt^n}x(t)$	$(j2\pi k f_0)^n \alpha_k$
Integration	$\int_T x(\tau)\,d\tau$	$(j2\pi k f_0)^{-1}\alpha_k, \quad \alpha_0 = 0$

where c_0 is the dc component, c_k and θ_k represent the amplitude and phase angle of the kth harmonic, respectively. Equation (23.22) is called the harmonic form of Fourier series expansion of $x(t)$. The parameters c_k and θ_k are related to a_k and b_k according to

$$c_0 = \frac{a_0}{2}, \qquad c_k = \sqrt{a_k^2 + b_k^2}, \qquad \theta_k = \tan^{-1}\frac{b_k}{a_k}$$

Properties of the Fourier Series[1,4]

Knowledge of signal symmetry can simplify its complex Fourier series coefficients computation. While many forms of symmetry can be established, the following important types of symmetry are more often encountered in signal analysis:

- even symmetry, $x(t) = x(-t)$
- odd symmetry, $x(t) = -x(t)$
- half-wave odd symmetry, $x(t) = -x(t + T/2)$

The effects of symmetry on the Fourier series computations are shown in Table 23.3. The other properties on Fourier series are summarized in Table 23.4, where α_k and β_k are the complex Fourier series coefficients of $x(t)$ and $y(t)$, respectively.

Fourier Transform

Frequency domain method of analyzing periodic CT signals has been presented in the previous section. A different technique, termed the Fourier transform, is used for the analysis of aperiodic signals. The development of the Fourier transform is based on Eqs. (23.16) and (23.17). Substituting (23.17) into (23.16) gives

$$x(t) = \sum_{k=-\infty}^{\infty}\left(\frac{1}{T}\int_{-T/2}^{T/2} x(\lambda)e^{-j2\pi k f_0 \lambda} d\lambda\right)e^{j2\pi k f_0 t} \tag{23.23}$$

Allowing $T \to \infty$, then $1/T \to df$, $kf_0 \to f$, and Eq. (23.23) becomes

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-j2\pi k f t} dt \tag{23.24a}$$

and

$$x(t) = \int_{-\infty}^{\infty} X(f)e^{j2\pi f t} df \tag{23.24b}$$

Properties of the Fourier Transform[5,6]

Some of the basic properties of the Fourier transform that are often used in analysis are given in Table 23.5.

The Fourier transform pairs of the following basic signals are also useful for analysis:

$$K \leftrightarrow K\delta(f)$$

$$\operatorname{sgn}(t) \leftrightarrow \frac{1}{j\pi f}$$

$$u(t) \leftrightarrow \frac{1}{2}\delta(f) + \frac{1}{j2\pi f}$$

$$\cos(2\pi f_0 t) \leftrightarrow \frac{1}{2}\delta(f - f_0) + \frac{1}{2}\delta(f + f_0)$$

A two-step procedure is required in order to determine the Fourier transform of a periodic signal. If $x(t)$ is a periodic signal with period T, then $x(t)$ can be Fourier series expanded as

$$x(t) = \sum_{k=-\infty}^{\infty} c_k e^{j2\pi k f_0 t}, \quad f_0 = \frac{1}{T} \tag{23.25}$$

Applying the linearity property to this equation gives

$$X(f) = \sum_{k=-\infty}^{\infty} c_k \delta(f - kf_0) \tag{23.26}$$

as the Fourier transform of any periodic signal $x(t)$. Eq. (23.26) explains the difference between the frequency spectrum produced by the Fourier series analysis and the amplitude spectral density produced by the Fourier transformation. Thus, the frequency spectrum is a (discrete) display of c_k versus kf_0, whereas

TABLE 23.5 Properties of the Fourier Transform

Property	Signal Description	Fourier Transform
Linearity	$ax(t) + by(t)$; a, b constants	$aX(f) + bY(f)$
Evenness and oddness	$x(-t) = x(t)$ $x(-t) = -x(t)$	$X(f) = 2\int_0^{\infty} x(t)\cos(2\pi ft)dt$ $X(f) = -2\int_0^{\infty} x(t)\sin(2\pi ft)dt$
Time shift	$x(t-\tau)$	$e^{-j2\pi f\tau}X(f)$
Time scale	$x(at)$	$\frac{1}{\lvert a\rvert}X\left(\frac{f}{a}\right)$
Time reversal	$x(-t)$	$X^*(f)$
Duality	$X(t)$	$x(-f)$
Time convolution	$x(t) * y(t)$	$X(f)\,Y(f)$
Frequency convolution	$x(t)\,y(t)$	$X(f) * Y(f)$
Modulation	$x(t)e^{j2\pi f_0 t}$	$X(f-f_0)$
Time differentiation	$\frac{d^n}{dt^n}x(t)$	$(j2\pi f)^n X(f)$
Frequency differentiation	$t^n x(t)$	$\left(\frac{j}{2\pi}\right)^n \frac{d^n}{df^n}X(f)$
Integration	$\int_{-\infty}^{\infty} x(\tau)d\tau$	$\frac{1}{j2\pi f}X(f) + \frac{1}{2}X(0)\delta(f)$
Correlation	$R_{xy}(\tau) = \int_{-\infty}^{\infty} y(t)x(t-\tau)dt$	$Y(-f)\,X(f)$
Parseval's theorem	$\int_{-\infty}^{\infty} \lvert x(t)\rvert^2 dt$	$\int_{-\infty}^{\infty} \lvert X(f)\rvert^2 df$

the amplitude spectral density gives a continuous display of the amplitude density spectrum, which in this case is in the form of impulses, rather than just a number.

Similarly the Fourier transform of a train of impulses of the form

$$p(t) = \sum_{n=-\infty}^{\infty} \delta(t-nT) \tag{23.27}$$

is given by

$$P(f) = \frac{1}{T}\sum_{k=-\infty}^{\infty} \delta(f-kF_s), \quad F_s = \frac{1}{T} \tag{23.28}$$

Energy and Power Spectral Density[6]

Suppose $x(t)$ is an aperiodic signal with a Fourier transform $X(f)$, then its energy is given by

$$E = R_{xx}(0) = \int_{-\infty}^{\infty} |x(t)|^2 dt = \int_{-\infty}^{\infty} |X(f)|^2 df \tag{23.29}$$

This is Parseval's theorem and it shows that the principle of conservation of energy in the time and frequency domains holds. The amplitude spectrum $X(f)$ can be expressed as

$$X(f) = |X(f)|\angle X(f)$$

and denoting

$$S_{xx}(f) = |X(f)|^2$$

then the total energy of the signal is given by

$$E = \int_{-\infty}^{\infty} S_{xx}(f)df \tag{23.30}$$

where $S_{xx}(f)$ represents the distribution of the signal energy as a function of frequency. $S_{xx}(f)$ is termed the energy spectral density for the finite energy signal $x(t)$.

Consider a periodic signal $x(t)$ with an autocorrelation function

$$R_{xx}(\tau) = \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} x(t)x^*(t-\tau)dt$$

Then,

$$R_{xx}(0) = \lim_{T\to\infty} \frac{1}{T}\int_{-T/2}^{T/2} |x(t)|^2 dt \tag{23.31}$$

which equals the signal power. Following the same procedure for energy signals, we define $S_{xx}(f)$ as the Fourier transform of $R_{xx}(\tau)$ so that

$$P = R_{xx}(0) = \int_{-\infty}^{\infty} S_{xx}(f)df \tag{23.32}$$

and $S_{xx}(f)$ is termed the power spectral density of the periodic signal $x(t)$.

The need to analyze stationary random signals also arises in many practical situations. The properties of such signals can be inferred from their correlation functions. For example the autocorrelation function, $\phi_{xx}(\tau)$ of a stationary random signal decreases and goes to zero as τ increases since the events become uncorrelated for a large separation of time. Hence, $\phi_{xx}(\tau) = \phi_{xx}(-\tau)$ and its Fourier transform exists. Consequently we can write

$$\phi_{xx}(0) = \int_{-\infty}^{\infty} \Gamma_{xx}(f)df \tag{23.33}$$

where $\Gamma_{xx}(f)$ and $\phi_{xx}(0)$ represent, respectively, the power spectral density and the average power of a random process.

Sampled Continuous-Time Signals

Discrete-time (DT) signals arise either naturally or by sampling continuous-time (CT) signals; however, the latter form is more often encountered in practice. In this case, a digital signal is formed from a CT signal through the process of analog-to-digital conversion. The first part of this process is the sampling of the analog signal, that is, the conversion of $x(t)$ into $x(nT_s)$, where T_s is the sampling period and its reciprocal, $F_s = 1/T_s$, is the sampling frequency in samples per second. The sampling frequency must be appropriately selected to avoid spectral distortion (aliasing), thereby ensuring that $x(t)$ can be reconstructed from its samples. To gain a good understanding of this procedure the sampling process is examined in the frequency domain.

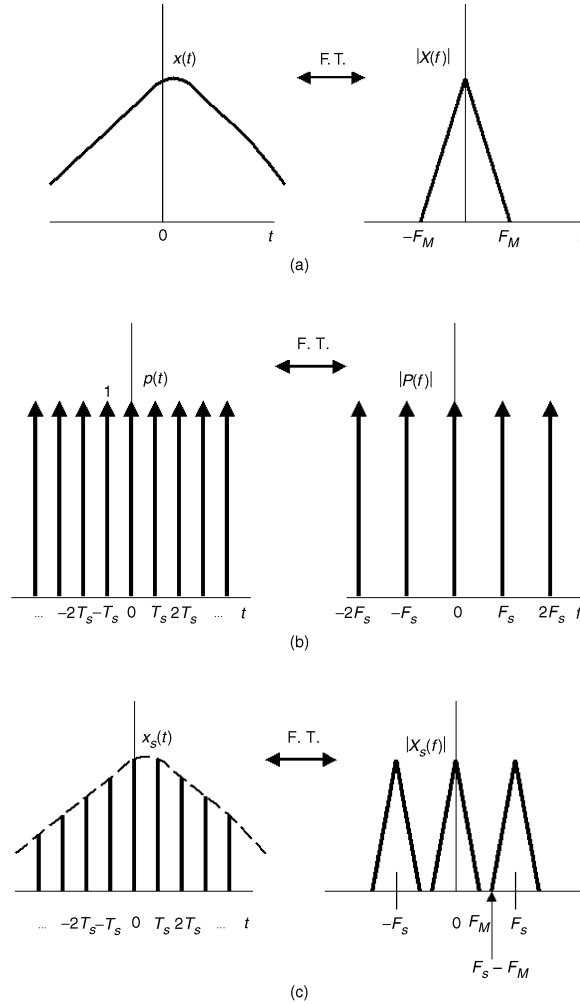



FIGURE 23.9 Ideal impulse sampling process: (a) bandlimited continuous-time signal and its Fourier transform, (b) train of impulses and its spectra, (c) sampled signal and its spectra.

Impulse Sampling[6–9]

Consider an idealized impulse-sampling process shown in Fig. 23.9, where $x(t)$ is to be sampled using the train of impulses $p(t)$, so that

$$x_s(t) = x(t)p(t) = x(t)\sum_{n=-\infty}^{\infty}\delta(t-nT_s) = \sum_{n=-\infty}^{\infty}x(nT_s)\delta(t-nT_s) \tag{23.34}$$

The main difference between $x_s(t)$ and $x(nT_s)$ is that the former is essentially a CT signal with zero values except at the integer values of T_s, while the latter is a perfectly selected sample of $x(t)$ as a result of impulse sampling.

The Fourier transform of Eq. (23.34) yields

$$X_s(f) = X(f) * P(f) = F_s\sum_{k=-\infty}^{\infty}X(f-F_s) \tag{23.33}$$

where $X(f)$ is the spectrum of $x(t)$.

It is observed that $X_s(f)$ consists of a scaled and periodically repeated version of $X(f)$ with period F_s. As shown in Fig. 23.9(c), it is obvious that when $F_s - F_M \geq F_M$, there is no overlapping of the spectra and the signal $x(t)$ can be recovered completely from $x_s(t)$. However, if $F_s - F_M < F_M$ the replicas of $X(f)$ will overlap, resulting in a distorted spectrum and as such, $x(t)$ can no longer be recovered from its sampled version. Consequently, in order to recover $x(t)$ from its samples, the sampling frequency should be such that

$$F_s - F_M \geq F_M$$

that is,

$$F_s \geq 2F_M$$

This is called the Nyquist sampling theorem. The minimum sampling frequency $F_s = 2F_M$ is called the Nyquist frequency. Sampling a signal at less than the Nyquist frequency results in a spectral distortion termed aliasing. Furthermore, sampling a signal at a frequency of at least Nyquist frequency implies that an ideal low-pass filter (LPF) with a gain of $1/F_s$ and cutoff frequency F_c can be used to recover its original spectrum, where $F_M \leq F_c \leq F_s - F_M$.

Suppose we want to reconstruct $x(t)$ from its samples. Assume that $X(f)$ is the spectrum of $x(nT_s)$, with no aliasing, as shown in Fig. 23.9(c). Thus,

$$X_a(f) = \begin{cases} \frac{1}{F_s} X(f), & |f| \leq \frac{F_s}{2} \\ 0, & |f| > \frac{F_s}{2} \end{cases} \tag{23.36}$$

Note that

$$X_a(f) = \sum_{n=-\infty}^{\infty} x(nT_s) e^{-j2\pi nfT_s} \tag{23.37}$$

so that its inverse Fourier transform is given by

$$\begin{aligned} x_a(t) &= \frac{1}{F_s} \sum_{n=-\infty}^{\infty} x(nT_s) \int_{-F_s/2}^{F_s/2} e^{j2\pi f(t-nT_s)} df \\ &= \sum_{n=-\infty}^{\infty} x(nT_s) \frac{\sin[\pi(t-nT_s)/T_s]}{\pi(t-nT_s)/T_s} \end{aligned} \tag{23.38}$$

This is the formula for the reconstruction of $x(t)$ from its samples. That is, $x(t)$ is generated by multiplying the appropriately shifted function $g(t) = \mathrm{sinc}(tF_s)$ by the corresponding samples of $x(nT_s)$.

Practical Sampling[8–10]

The above discussion on sampling is based on the idealized models of periodic impulse sampling and bandlimited interpolation. In practice, CT signals are not precisely bandlimited just as impulse signals and ideal low-pass filters do not exist physically. Figure 23.10 represents the block diagram for the conversion of continuous signals into their discrete forms. The continuous-time signal is prefiltered, sampled, quantized, and finally encoded into finite-length words of, say, b bits. The prefilter, which is also called anti-aliasing filter (AAF), is a low-pass filter that is needed to limit the input signal bandwidth to $F_s/2$ prior to sampling to avoid aliasing. In practice this filter will possess non-ideal characteristics, hence

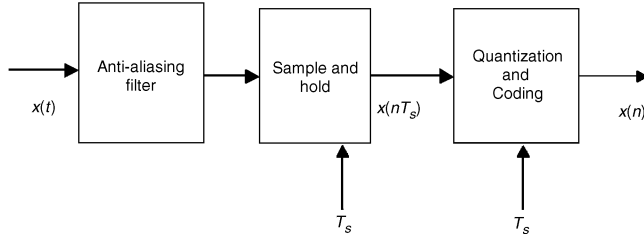



FIGURE 23.10 Practical sampling of continuous-time signals.

it should be designed to provide sufficient attenuation, usually to a level undetectable by the analog-to-digital converter (ADC) at frequencies above the Nyquist frequency.

The prefiltered signal is fed into the ADC where it will be converted to a DT signal. The ADC has an in-built sample-and-hold circuit and it operates at the sampling rate of F_s; however, the sampling function has a finite width as opposed to the impulse sampling discussed above. The sampling operation can be modeled by the finite-width pulse sampler shown in Fig. 23.11(b), in which the sampling gate is open for τ-out of T_s seconds and shorted to ground the remainder of the sampling interval. Here $p(t)$ is expressed as

$$p(t) = \sum_{n=-\infty}^{\infty} \prod\left(\frac{t-nT_s}{\tau}\right)$$

which can be Fourier series expanded as

$$p(t) = \sum_{n=-\infty}^{\infty} c_k e^{j2\pi k f_s t}$$

where

$$c_k = \frac{\tau}{T_s}\sin c\left(k\frac{\tau}{T_s}\right)$$

The Fourier transform of the sampled signal can be written as

$$X_s(f) = \sum_{n=-\infty}^{\infty} c_k X(f-kF_s)$$

Note that c_k is not constant in this expression (as opposed to the impulse sampling) since its value depends on the harmonic number (k) as well as the duty cycle τ/T_s. The discrete-time signal, $x(nT_s)$, is fed into the quantizer where each sample is transformed into one of the nearest finite sets of prescribed values, that is, $\hat{x}(n) = Q(x(nT_s))$, where $\hat{x}(n)$ is the quantized sample. The quantization process is shown in Fig. 23.11(d) for zero-order sample hold ADC, where L_i denotes the quantization level and Δ is the quantization step. The quantization error (or noise) incurred in this process is

$$L_i - \frac{\Delta}{2} < x(nT_s) < L_i + \frac{\Delta}{2}$$

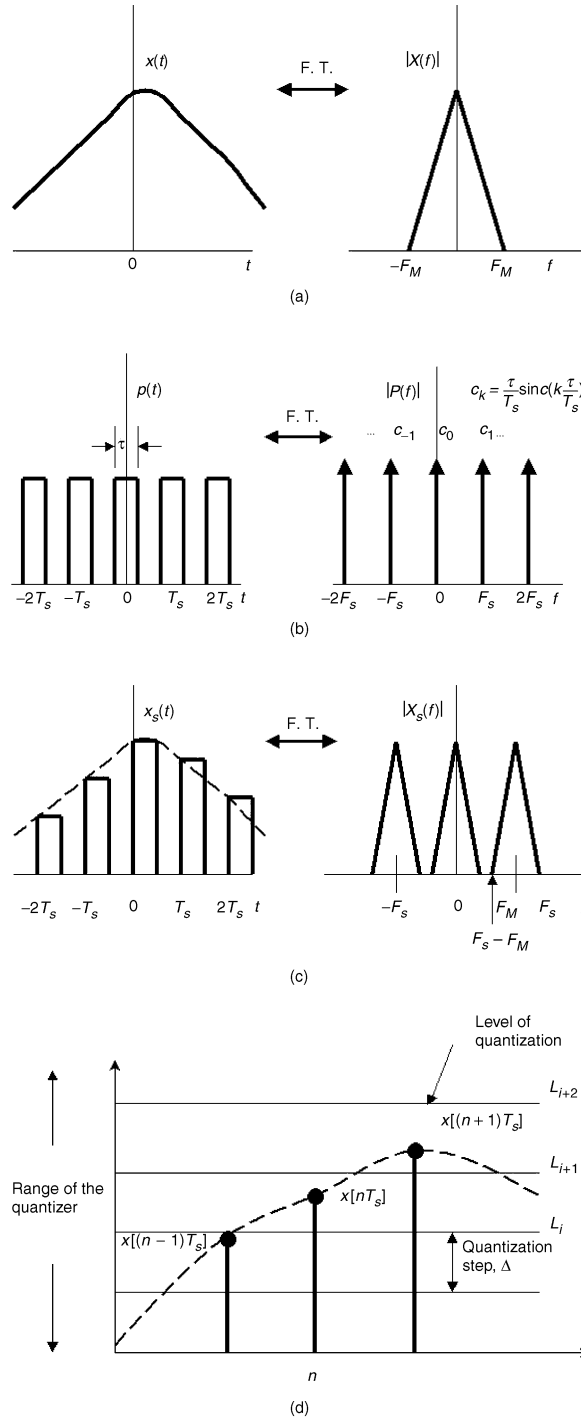



FIGURE 23.11 Finite-width pulse sampling signals and spectra: (a) bandlimited signal and spectrum, (b) finite-width train of pulses and its transform, (c) sampled signal and its spectra, (d) quantization process.

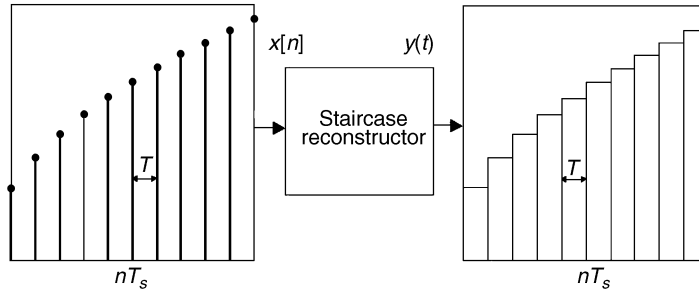



FIGURE 23.12 Digital-to-analog conversion process.

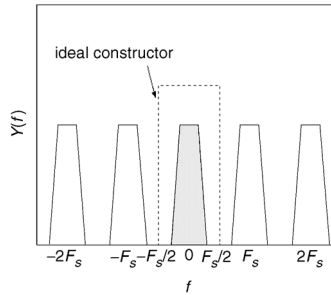



FIGURE 23.13 Frequency response of the ideal reconstruction filter.

From statistical considerations, the noise power is found to be $\Delta^2/12$ watts. A common measure of the performance of the ADC is the ratio of the signal power to noise power, and this is called the signal-to-quantization noise power, which expressed in decibels (dB) is

$$\mathrm{SQNR(dB)} = 1.76 + 6.02b$$

Digital-to-Analog Conversion[8–12]

Reconstruction of the analog signal from its sampled form is closely akin to lowpass filtering of the sampled signal. Figure 23.12 shows how an analog signal can be reconstructed by filling the gaps between samples and holding the current value constant till the next sample is received. Consider an ideal reconstruction filter with an impulse response function $h(t)$, then its response is given by

$$y(t) = \sum_{n=-\infty}^{\infty} x(nT_s)h(t - nT_s)$$

Taking the Fourier transform of this equation gives

$$Y(f) = H(f)X(f)$$

where $X(f)$ is the periodic spectrum of $x(nT_s)$ as shown in Fig. 23.13 and $h(t) = \sin c(F_s t)$. Note that $h(t)$ is noncausal; hence it cannot be used for real-time applications. Furthermore, since $h(t)$ is not time-limited, an infinite number of impulse responses must be used for interpolating between values in order to obtain exact results. Consequently, alternative reconstruction filters such as zero-order hold (staircase), first-order, or fractional-order holds are used in practice. However, the staircase reconstruction filter is, by and large, the simplest and most widely used in practice. The impulse response of this filter is given

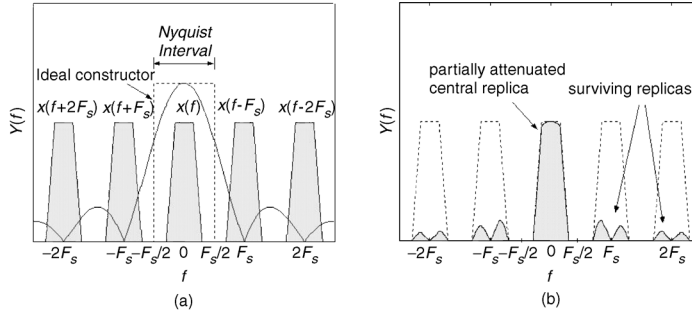



FIGURE 23.14 (a) Staircase reconstruction in frequency domain, (b) Result of staircase reconstruction in frequency domain.

as $h_o(t) = u(t) - u(t - T_s)$ since it must have a duration of T_s seconds to fill the entire gap between the samples. Thus, in effect this filter, as its name implies, generates a staircase approximation to the original analog signal. The frequency response of the filter is

$$H_o(f) = T\frac{\sin(\pi f T_s)}{\pi f T_s}e^{-j\pi f T_s}$$

Though the output of the staircase reconstruction filter is smoother than its sampled form, see Fig. 23.14(a), it contains spurious high-frequency components due to the sudden jumps in the staircase levels as different sampled values are considered. In addition, holding each of $x(nT_s)$ by T_s seconds introduces a time shift of $T_s/2$ to the output signal. However, this time delay has virtually no effect on the quality of the output signal. Figure 23.14(b) compares the signal spectra before and after the staircase reconstruction filter. It is noted that the output spectrum is slightly distorted due to non-ideal characteristics of $H_o(f)$ and that distorted and attenuated versions of $X(f)$ remain centered at nonzero multiples of F_s. These remaining spectral replicas may be removed by using an anti-imaging filter.[8,11,12] In essence, the anti-imaging filter smoothens out the discontinuities produced by the staircase reconstruction filter as illustrated in Fig. 23.15.

Frequency Analysis of Discrete-Time Signals

The analysis of discrete-time (DT) signals in the frequency domain is very much similar to that of continuous-time (CT) signals. As in the CT analysis, the techniques for the analysis depend on the type of signal. Analysis of aperiodic DT signals will be considered first.

Discrete-Time Fourier Transform[6–8]

The decomposition of an aperiodic DT signal into its frequency components is carried out using discrete-time Fourier transformation (DTFT). Thus, the DTFT of $x(n)$ is given by

$$X(f) = \sum_{n=-\infty}^{\infty} x(n)e^{-j2\pi fn} \tag{23.39}$$

Unlike the Fourier transform of analog signal, $X(f)$ is periodic with period F_s; hence, the frequency range for a DT signal is unique over the frequency interval $(-F_s/2, F_s/2)$ or, equivalently $(0, F_s)$. Note that Eq. (23.39) must be absolutely summable in order for $X(f)$ to exist, that is,

$$\sum_{n=-\infty}^{\infty} |x(n)| < \infty$$

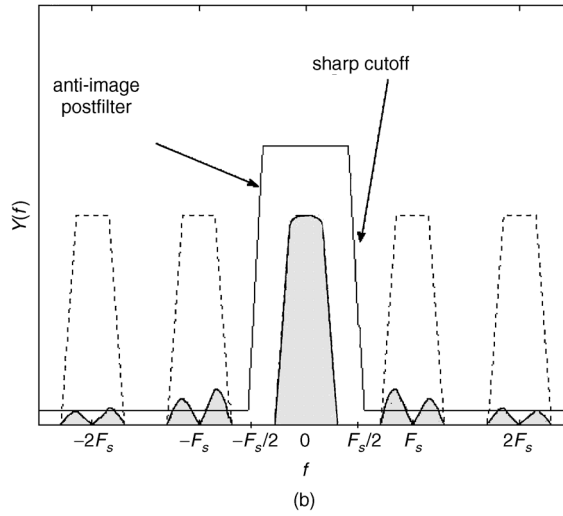
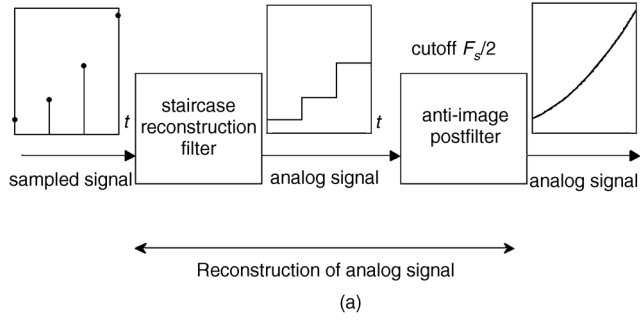



FIGURE 23.15 (a) Reconstruction of analog signal, (b) effect of anti-image postfiltering.

If the spectrum of a signal exists, then we can find the signal from its spectrum through the inverse DTFT. The inverse DTFT of $X(f)$ is given as

$$x(n) = \frac{1}{F_s}\int_{-F_s/2}^{F_s/2} X(f)e^{j2\pi fn}df$$

The properties of the DTFT of a DT signal are shown in Table 23.6.

For DT signals, the convolution of two sequences $x(n)$ and $y(n)$ is expressed as

$$x(n) * y(n) = \sum_{m=-\infty}^{\infty} x(m)y(n-m) \tag{23.40}$$

while the cross-correlation function of $x(n)$ and $y(n)$ is given by

$$R_{xy}(n) = x(n) \oplus y(n) = \sum_{m=-\infty}^{\infty} x(m)y^*(m-n)$$

When $x(n) = y(n)$, then the autocorrelation of $x(n)$ is

$$R_{xx}(n) = \sum_{m=-\infty}^{\infty} x(m)x^*(m-n)$$

TABLE 23.6 Properties of DTFT

Property	Signal Description	Frequency Domain
Even symmetry (real signal)	$x_e(n) = \frac{1}{2}\{x(n)+x(-n)\}$	$X_e(f) = \sum_{n=-\infty}^{\infty} x_e(n)\cos(2\pi nf)$
Odd symmetry (real signal)	$x_o(n) = \frac{1}{2}\{x(n)-x(-n)\}$	$X_o(f) = -\sum_{n=-\infty}^{\infty} x_o(n)\sin(2\pi nf)$
Linearity	$ax(n) + by(n)$	$aX(f) + bY(f)$
Time shifting	$x(n-m)$	$e^{-j2\pi fm}X(f)$
Time reversal	$x(-n)$	$X(-f)$
Convolution	$x(n)*y(n)$	$X(f)Y(f)$
Correlation	$R_{xy}(n) = x(n) \oplus y(n)$	$S_{xy}(f) = X(f)Y(-f)$
Wiener–Khintchine theorem	$R_{xx}(n)$	$S_{xx}(f)$
Frequency shifting	$e^{j2\pi f_0 n}x(n)$	$X(f-f_0)$
Modulation	$x(n)\cos(2\pi nf_0)$	$\frac{1}{2}\{X(f+f_0)+X(f-f_0)\}$
Multiplication	$x(n)y(n)$	$\frac{1}{F_s}\int_{F_s} X(\lambda)Y(f-\lambda)d\lambda$
Differentiation in the frequency domain	$nx(n)$	$\frac{j}{2\pi}\frac{dX(f)}{df}$
Time differencing	$x(n)-x(n-1)$	$(1-e^{-j2\pi f})X(f)$
Summation	$\sum_{m=-\infty}^{n} x(m)$	$\frac{X(f)}{(1-e^{-j2\pi f})} + \frac{F_sX(0)}{2}\sum_{m=-\infty}^{\infty}\delta(f-mF_s)$
Conjugation	$x^*(n)$	$X^*(-f)$
Parseval's theorem	$\sum_{n=-\infty}^{\infty} x(n)y^*(n)$	$\frac{1}{F_s}\int_{F_s} X(f)Y^*(f)df$

With the exception of the reflection operation (for the convolution), the procedures for computing the convolution and correlation are the same. Hence, it is more computationally efficient to use the same algorithm for evaluating both functions. To achieve this, one of the sequences is reflected (only for the correlation analysis), followed by convolution operation, that is,

$$R_{xy}(n) = x(n) * y^*(-n)$$

and

$$R_{xx}(n) = x(n) * x^*(-n)$$

The energy of an aperiodic signal is computed from

$$E = \sum_{n=-\infty}^{\infty} \left|x(n)^2\right| = \sum_{n=-\infty}^{\infty} x(n)x^*(n)$$

Substituting the conjugated form of (23.40) into this equation gives

$$E = \frac{1}{F_s}\int_{F_s} |X(f)|^2 df = \frac{1}{F_s}\int_{F_s} S_{xx}(f)df \tag{23.41}$$

This expression relates the distribution of the energy of an aperiodic signal to frequency. The quantity $|X(f)|^2$ is called the energy spectral density of $x(n)$. The DTFT of some commonly encountered signals are shown in Table 23.7.

TABLE 23.7 DTFT of Common DT Signals

$x(n)$	Frequency-Domain Representation, $X(f)$
$\delta(n)$	1
$A, -\infty < n < \infty$	$AF_s \sum_{k=-\infty}^{\infty} \delta(f - mF_s), \; F_s = \frac{1}{T}$
$u(n)$	$\frac{1}{1 - e^{-j2\pi f}} + \frac{F_s}{2} \sum_{k=-\infty}^{\infty} \delta(f - kF_s)$
$\Pi\left(\frac{n}{2q+1}\right)$	$\frac{\sin[(2q+1)\pi f]}{\sin(\pi f)}$
$\Lambda\left(\frac{n}{q}\right)$	$\frac{\sin^2(\pi f q)}{q \sin^2(\pi f)}$
$\mathrm{sgn}(n)$	$\frac{2}{1 - e^{-j2\pi f}}$
$\alpha^n u(n)$	$\frac{1}{1 - \alpha e^{-j2\pi f}}, \; \lvert\alpha\rvert < 1$
$\alpha^{\lvert n\rvert}$	$\frac{1 - \alpha^2}{1 - 2\alpha\cos(2\pi f) + \alpha^2}, \; \lvert\alpha\rvert < 1$
$n\alpha^n u(n)$	$\frac{\alpha e^{-j2\pi f}}{(1 - \alpha e^{-j2\pi f})^2}, \; \lvert\alpha\rvert < 1$
$e^{-\alpha n} u(n)$	$\frac{1}{1 - e^{-(\alpha + j2\pi f)}}, \; \lvert\alpha\rvert > 0$
$e^{-\alpha\lvert n\rvert}$	$\frac{1 - e^{-2\alpha}}{1 - 2e^{-\alpha}\cos(2\pi f) + e^{-2\alpha}}, \; \lvert\alpha\rvert > 0$
$e^{j2\pi f_0 n}$	$F_s \sum_{k=-\infty}^{\infty} \delta(f - f_0 + kF_s)$
$\cos(2\pi f_0 n + \theta)$	$\frac{F_s}{2} \sum_{k=-\infty}^{\infty} \{e^{j\theta}\delta(f - f_0 + kF_s) + e^{-j\theta}\delta(f + f_0 + kF_s)\}$
$\frac{\sin(2\pi f_c n)}{\pi n}$	$\Pi\left(\frac{f}{2f_c}\right)$

Discrete Fourier Series[6–8]

Suppose $x(n)$ is a periodic DT signal with period N, then it is possible to obtain its discrete Fourier series (DFS) expansion in a manner analogous to the computation of the complex Fourier series for the CT signals. The orthogonal basis function for the DFS is $W_N^{-n} = e^{j2\pi n/N}$ so that the decomposition of $x(n)$ into a sum of N harmonically related complex exponentials is expressed as

$$x(n) = \sum_{k=0}^{N-1} c_k e^{j2\pi kn/N} = \sum_{k=0}^{N-1} c_k W_N^{-kn} \tag{23.42}$$

where c_k represents the discrete Fourier series coefficients. Multiplying both sides of (23.42) by W_N^{mn}, summing over one period, and using the fact that

$$\sum_{k=0}^{N-1} W_N^{n(k-m)} = \begin{cases} N, & k = m \\ 0, & k \neq m \end{cases}$$

TABLE 23.8 Properties of the Discrete Fourier Series

Property	Signal Description	Discrete Fourier Series Coefficients
Linearity	$\sum_{m=0}^{Q} \beta_m x_m(n)$	$\sum_{m=0}^{Q} \beta_m c_{m,k}$
Time shift	$x(n-m)$	$c_k W_N^{km}$
Time-reversal	$x(-n)$	c_{-k}
Modulation	$x(n) W_N^{-mn}$	c_{k-m}
Real $x(n)$	$x(n)$	$c_k = c_k^*$
$x_e(n)$ is an even DT signal	$x_e(n)$	$\mathrm{Re}(c_k)$
$x_0(n)$ is an odd DT signal	$x_o(n)$	$j\{\mathrm{Im}(c_k)\}$

results in

$$c_k = \frac{1}{N}\sum_{n=0}^{N-1} x(n) W_N^{kn}, \quad k = 0, 1, 2, \ldots, N-1 \tag{23.43}$$

The DFS coefficients $\{c_k\}$ provide the description of $x(n)$ in the frequency domain since c_k represents the amplitude and phase spectra associated with the kth harmonic component. Note that $\{c_k\}$ is a periodic sequence with fundamental period N, that is, $c_k = c_{k+N}$. Hence, any N consecutive samples of the signal or its DFS coefficients provide a complete description of the signal in the time or frequency domains. As shown in Table 23.8, the properties of the DFS follow closely those given for the DTFT. One of the main advantages of DFS over the DTFT is the replacement of the infinite summation in the DTFT by a finite sum in the DFS, thus allowing the computations of DFS and inverse DFS by digital computers.

Consider a periodic DT signal with period N, then its average power is

$$P = \frac{1}{N}\sum_{n=0}^{N-1} |x(n)|^2 \tag{23.44}$$

Using Eq. (23.42) in (23.44) gives

$$P = \frac{1}{N}\sum_{n=0}^{N-1} x(n)x^*(n) = \frac{1}{N}\sum_{n=0}^{N-1} x(n)\left(\sum_{k=0}^{N-1} c_k^* W_N^{kn}\right) = \sum_{k=0}^{N-1} |c_k|^2 \tag{23.45}$$

This is a Parseval's relation for the DT periodic signals, which indicates that the average power in $x(n)$ is the sum of the harmonics power. Consequently, the sequence $|c_k|^2$ is the power spectral density as this represents the distribution of power as a function of frequency.

The Discrete Fourier Transform[6,8,13]

The discrete Fourier transform (DFT) is an important and extremely powerful technique for analyzing DT signals. Contrary to the DTFT, the DFT is applicable to finite-length sequences and it produces finite-length discrete spectra. Consequently, this transformation is amenable to digital computations and it is suitable for use in digital hardware implementations. As a result of its practicability, DFT has become a very useful tool for analyzing various waveforms or data that arise in many disciplines.

The DFT is a mapping of an N sample sequence, $x(n)$, into another N sequence $X(k)$ in the frequency domain, that is,

$$X(k) = \sum_{n=0}^{N-1} x(n) W_N^{kn}, \quad k = 0, 1, \ldots, N-1 \tag{23.46}$$

TABLE 23.9 Properties of DFT

Property	Signal Description	Discrete Fourier Transform
Linearity	$\sum_{m=0}^{Q} a_m x_m(n)$	$\sum_{m=0}^{Q} a_m X_m(k)$
Circular shift	$x[(n-m) \bmod N]$	$W_N^{km} X(k)$
Modulation	$W_N^{-qn} x(n)$	$X[(k-q) \bmod N]$
Time reversal	$x^*[-n \bmod N]$	$x^*(k)$
Complex conjugation	$x^*(n)$	$X^*[-k \bmod N]$
Circular convolution	$\sum_{m=0}^{N-1} x(m) y[(n-m) \bmod N]$	$X(k)Y(k)$
Multiplication	$x(n)y(n)$	$\frac{1}{N}\sum_{m=0}^{N-1} X(m) Y[(k-m) \bmod N]$
Parseval's theorem	$\sum_{k=0}^{N-1} x(n) y^*(n)$	$\frac{1}{N}\sum_{k=0}^{N-1} X(k) Y^*(k)$
Real part of signal	$\mathrm{Re}\{x(n)\}$	$\frac{1}{2}\{X(k) + X^*(N-k)\}$
Imaginary part of signal	$j\,\mathrm{Im}\{x(n)\}$	$\frac{1}{2}\{X(k) - X^*(N-k)\}$
Complex even	$x_{ce}(n) = \frac{1}{2}\{x(n) + x^*(N-n)\}$	$X_R(k)$
Complex odd	$x_{co}(n) = \frac{1}{2}\{x(n) - x^*(N-n)\}$	$jX_I(k)$
Any real signal	$x(n)$	$X(k) = X^*(N-k)$ $X_R(k) = X_R(N-k)$ $X_I(k) = -X_I(N-k)$ $\lvert X(k)\rvert = \lvert X(N-k)\rvert$ $\angle X(k) = -\angle X(N-k)$

$X(k)$ is called the kth harmonic and this exists provided all the samples of $x(n)$ are bounded. In order to recover $x(n)$ from $X(k)$ we need to perform inverse transformation. Thus, the inverse discrete Fourier transform (IDFT) of $X(k)$ is defined as

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(k) W_N^{-kn}, \quad n = 0, 1, \ldots, N-1 \tag{23.47}$$

The DFT and DFS are conceptually different in the sense that the DFS is only applicable to periodic signals whereas DFT assumes that the signal is periodic, that is, there is an inherent windowing of nonperiodic signal if analyzed by the DFT technique. Also the scaling factor is applicable to the synthesis equation for the DFT operation, whereas it is used in the analysis equation in the DFS analysis.

The DFT properties bear strong resemblance to those of the DFS and DTFT as shown in Table 23.9.

The following should be noted in the DFT computation: Circular shift operation can be considered as wrapping the part of the sequence that falls outside of 0 to $N-1$ to the front of the sequence, that is, $x[(-n) \bmod N]$ is equivalent to $x(N-n)$.

As a result of the circular shift, linear convolution as given by Eq. (23.40) is different from circular convolution given in Table 23.9. Thus,

$$x(n) \otimes_N y(n) = \sum_{m=0}^{N-1} x(m) y[(n-m) \bmod N] = \sum_{m=0}^{N-1} x[(n-m) \bmod N] y(m)$$

where $x[(n-m) \bmod N]$ is the reflected and circularly translated version of $x(n)$. However, by appropriate selecting the value of N, both the circular and linear convolution can be the same. Thus, if signals $x(n)$ and $y(n)$ are of length N_1 and N_2, respectively, then circular and linear convolution are the same provided $N \geq N_1 + N_2 - 1$.

Circular correlation can be implemented by circular convolution since the cross correlation of the two sequences $x(n)$ and $y(n)$ can be expressed as

$$R_{xy}(n) = x(n) \otimes_N y(n) = x(n) \otimes_N y^*[(-n) \bmod N]$$

From Eqs. (23.46) and (23.47), it is noted that to compute the DFT coefficients would require N^2 complex multiplications and $N(N-1)$ complex additions. This can be a computational load if N is very large. Fast Fourier transforms (FFTs) are different types of algorithms that have been developed to speed up the computation of the DFT coefficients. Readers can refer to many textbooks where the development of the various forms of the FFT algorithms has been discussed.[7,8] It can be shown that the number of computations required for the FFT algorithms can be expressed as a constant times $N * \log_2 N$. Consequently, there is a reduction in computation when an FFT algorithm is used for the DFT computation.

Remarks on the DFT Processing of Signals

Zero Padding[6,8]

The use of FFT for DFT coefficients computation imposes some constraints on the value of N, for example N has to be a power of 2 for radix-2 FFT algorithm. Also, circular convolution is an undesirable solution as it manifests from the IDFT of the product of the transform of two sequences. Zero-padding is a technique for remedying the above situations, that is, the zero-padding is used either to augment the sequence length so that either a radix-2 FFT algorithm can be used or to ensure that both the linear and circular convolution are the same. In addition, this procedure is also used to provide a better-looking display of the signal spectrum since the frequency spacing of the FFT samples decreases as N increases.

Error Sources[8,10]

The resultant spectrum of a sampled signal comprises the analog spectrum repeated at the integer multiples of the sampling frequency. The overlapping of the analog signal spectrum with its shifted versions causes *aliasing*. In practice, excessive errors due to aliasing are minimized by either increasing the sampling rate or prefiltering the signal to remove the high-frequency spectral components. Another source of error in the DFT processing of signals is the *spectral leakage*. This is caused by using a window to truncate an infinitely long signal to obtain a finite-length data for DFT processing. It is known that windowing the samples of a signal in the time domain transforms to convolution of sampled signal spectrum and the window spectrum in the frequency domain. Suppose the window width does not correspond to an integer multiple of periods of all the frequency components of a discrete-time signal, then a single frequency component will spread (leak) into other frequency locations in the DFT of the truncated data. This phenomenon causes spectral distortion and makes it difficult to determine whether or not two closely adjacent frequencies are present in a signal. Spectral resolution becomes better if the window width is increased, or by choosing a window function with low sidelobes. The *picket-fence effect* arises because only a finite number of frequency points of a continuous-frequency spectrum are produced by the DFT. It is therefore possible to miss the peak of a particular frequency component in a signal because it is located between two adjacent frequency points in the spectrum. Since the frequency spacing $\Delta f = F_s/N$, this problem can be alleviated by increasing N the number of DFT points while maintaining the same sampling rate, implying having more samples in the DFT or a employing zero-padding technique.

DFT Parameter Selection[6]

The sampling period, T_s, frequency resolution (spacing), Δf, and the DFT length, N, are the three parameters that must be specified in performing DFT signal processing. Since $F_s = N\Delta f = 1/T_s$, these parameters are related according to $\Delta f = 1/NT_s$. If T_x denotes the length of the sampled data, then $T_x = MT_s$. Thus, there is

no spectral distortion if $N \geq M = T_x/T_s$. Equivalently, there will be no spectral distortion if

$$\Delta f = \frac{1}{T} = \frac{1}{NT_s} \leq \frac{1}{MT_s} = \frac{1}{T_x}$$

Parameters for the DFT processing of a sampled continuous signal must be carefully selected to avoid spectral distortion due to aliasing or data truncation. Assuming a window width of T_x seconds and that the signal has a maximum bandwidth of B hertz, then based on the sampling theorem we would have negligible aliasing, provided $F_s = 1/T_s \geq 2B$. Spectral leakage due to sharp data truncation is avoided provided the frequency resolution is selected to satisfy $1/\Delta f = T \geq T_x$. Consequently, spectral distortion due to aliasing and spectral leakage can be avoided if the length of the DFT is selected to satisfy $N = F_s/\Delta f \geq 2BT_x$.

References

1. Soliman, S.S., and Srinath, M.D., *Continuous and Discrete Signals and Systems,* Prentice-Hall, 1998.
2. O'Flynn, M., and Moriarty, E., *Linear Systems: Time Domain and Transform Analysis,* Wiley, 1987.
3. Lathi, B.P., *Modern Digital and Analog Communication Systems,* Oxford University Press, Third Edition, 1998.
4. Proakis, J.G., and Salehi, M., *Communication Systems Engineering,* Prentice-Hall, 1994.
5. Taylor, F.J., *Principles of Signals and Systems,* McGraw-Hill, 1994.
6. Carlson, G.E., *Signal and Linear System Analysis,* Wiley, Second Edition, 1998.
7. Proakis, J.G., and Manolakis, D.G., *Digital Signal Processing: Principles, Algorithms, and Applications,* Prentice-Hall, 1996.
8. Oppenheim, A.V., and Schafer, R.W., with Buck, J.R., *Discrete-Time Signal Processing,* Prentice-Hall, Second Edition, 1999.
9. Houts, R.C., *Signal Analysis in Linear Systems,* Saunders College Publishing, 1991.
10. Ziemer, R.E., Tranter, W.H., and Fannin, D.R., *Signals and Systems: Continuous and Discrete,* MacMillan Publishing Company, Third Edition, 1993.
11. Orfanidis, S.J., *Introduction to Signal Processing,* Prentice-Hall, 1996.
12. Haykin, S., and Veen, B.V., *Signals and Systems,* Wiley, 1999.
13. Taylor, F., and Mellot, J., *Hands-On Digital Signal Processing,* McGraw-Hill, 1998.

23.2 *z* Transform and Digital Systems

Rolf Johansson

A digital system (or discrete-time system or sampled-data system) is a device such as a digital controller or a digital filter or, more generally, a system intended for digital computer implementation and usually with some periodic interaction with the environment and with a supporting methodology for analysis and design. Of particular importance for modeling and analysis are recurrent algorithms—for example, difference equations in input–output data—and the *z* transform is important for the solution of such problems.

The *z* **transform** is being used in the analysis of linear time-invariant systems and discrete time signals—for example, for digital control or filtering—and may be compared to the Laplace transform as used in the analysis of continuous-time signals and systems, a useful property being that the convolution of two time-domain signals is equivalent to multiplication of their corresponding *z* transforms. The *z* transform is important as a means to characterize a linear time-invariant system in terms of its pole–zero locations, its transfer function and Bode diagram, and its response to a large variety of signals. In addition, it provides important relationships between temporal and spectral properties of signals. The *z* transform generally appears in the analysis of difference equations as used in many branches of engineering and applied mathematics.

The z Transform

The z transform of the sequence $\{x_k\}_{-\infty}^{+\infty}$ is defined as the generating function

$$X(z) = \mathcal{Z}\{x\} = \sum_{k=-\infty}^{\infty} x_k z^{-k} \tag{23.48}$$

where the variable z has the essential interpretation of a forward shift operator so that

$$\mathcal{Z}\{x_{k+1}\} = z\mathcal{Z}\{x_k\} = zX(z) \tag{23.49}$$

The z transform is an infinite power series in the complex variable z^{-1} where $\{x_k\}$ constitutes a sequence of coefficients. As the z transform is an infinite power series, it exists only for those values of z for which this series converges and the *region of convergence* of $X(z)$ is the set of z for which $X(z)$ takes on a finite value. A sufficient condition for existence of the z transform is convergence of the power series

$$\sum_{k=-\infty}^{\infty} |x_k| \cdot |z^{-k}| < \infty \tag{23.50}$$

The region of convergence for a finite-duration signal is the entire z plane except $z = 0$ and $z = \infty$. For a one-sided infinite-duration sequence $\{x_k\}_{k=0}^{\infty}$, a number r can usually be found so that the power series converges for $|z| > r$. Then, the *inverse z transform* can be derived as

$$x_k = \frac{1}{2\pi i}\oint X(z) z^{k-1} dz \tag{23.51}$$

where the contour of integration encloses all singularities of $X(z)$. In practice, it is standard procedure to use tabulated results; some standard z transform pairs are to be found in Table 23.10.

Digital Systems and Discretized Data

Periodic sampling of signals and subsequent computation or storing of the results requires the computer to schedule sampling and to handle the resulting sequences of numbers. A measured variable $x(t)$ may be available only as periodic observations of $x(t)$ as sampled with a time interval T (the sampling period). The sample sequence can be represented as

$$\{x_k\}_{-\infty}^{\infty}, \quad x_k = x(kT) \quad \text{for } k = \ldots, -1, 0, 1, 2, \ldots \tag{23.52}$$

and it is important to ascertain that the sample sequence adequately represents the original variable $x(t)$; see Fig. 23.16. For ideal sampling it is required that the duration of each sampling be very short and the sampled function may be represented by a sequence of infinitely short impulses $\delta(t)$ (the Dirac impulse). Let the sampled function of time be expressed thus:

$$x_{\Delta}(t) = x(t) \cdot T\sum_{k=-\infty}^{\infty} \delta(t - kT) = x(t) \cdot \sqcup\!\sqcup_T(t) \tag{23.53}$$

where

$$\sqcup\!\sqcup_T(t) \stackrel{\Delta}{=} T\sum_{k=-\infty}^{\infty} \delta(t - kT) \tag{23.54}$$

TABLE 23.10 Properties of the z Transform

z transform	$\mathcal{Z}(\{f_k\}) = F(z)$
Convolution	$\mathcal{Z}(\{f_k * g_k\}) = \mathcal{Z}(\{f_k\}) \cdot \mathcal{Z}(\{g_k\})$
	$\mathcal{Z}(\{f_k \cdot g_k\}) = \mathcal{Z}(\{f_k\}) * \mathcal{Z}(\{g_k\})$
Forward shift	$\mathcal{Z}(\{f_{k+1}\}) = z\mathcal{Z}(\{f_k\}) = zF(z)$
Backward shift	$\mathcal{Z}(\{f_{k-1}\}) = \mathcal{Z}(\{f_k\}) = z^{-1}F(z)$
Linearity	$\mathcal{Z}(\{af_k + bg_k\}) = a\mathcal{Z}(\{f_k\}) + b\mathcal{Z}(\{g_k\})$
Multiplication	$\mathcal{Z}(\{a^k f_k\}) = F(a^{-1}z)$
Final value	$\lim_{k\to\infty} f_k = \lim_{k\to 1}(1-z^{-1})F(z)$
Initial value	$f_0 = \lim_{z\to\infty} F(z)$

	Time Domain		z Transform
Impulse	$\delta_k = \begin{cases} 1, & k = 0 \\ 0, & k \neq 0 \end{cases}$	$\leftrightharpoons$	$\mathcal{Z}\{\delta_k\} = 1, \quad z \in C$
Step function	$\sigma_k = \begin{cases} 0, & k < 0 \\ 1, & k \geq 0 \end{cases}$	$\leftrightharpoons$	$\mathcal{Z}(\sigma_k) = \frac{z}{z-1}, \quad \lvert z\rvert > 1$
Ramp function	$x_k = k \cdot \sigma_k$	$\leftrightharpoons$	$X(z) = \frac{z}{(z-1)^2}, \quad \lvert z\rvert > 1$
Exponential	$x_k = a^k \cdot \sigma_k$	$\leftrightharpoons$	$X(z) = \frac{z}{z-a}, \quad \lvert z\rvert > \lvert a\rvert$
Sinusoid	$x_k = \sin \omega k \cdot \sigma_k$	$\leftrightharpoons$	$X(z) = \frac{z \sin \omega}{z^2 - 2z\cos\omega + 1}, \quad \lvert z\rvert > 1$

FIGURE 23.16 A continuous-time signal $x(t)$ and a sampling device that produces a sample sequence $\{x_k\}$.

and where the sampling period T is multiplied to ensure that the averages over a sampling period of the original variable x and the sampled signal x_Δ, respectively, are of the same magnitude. A direct application of the discretized variable $x_\Delta(t)$ in Eq. (23.53) verifies that the spectrum of x_Δ is related to the z transform $X(z)$ as

$$X_\Delta(i\omega) = \mathcal{F}\{x(t) \cdot ⧢_T(t)\} = T\sum_{k=-\infty}^{\infty} x_k \exp(-i\omega kT) = TX(e^{i\omega T}) \tag{23.55}$$

Obviously, the original variable $x(t)$ and the sampled data are not identical, and thus it is necessary to consider the distortive effects of discretization. Consider the spectrum of the sampled signal $x_\Delta(t)$ obtained as the Fourier transform

$$X_\Delta(i\omega) = \mathcal{F}\{x_\Delta(t)\} = \mathcal{F}\{x(t)\} * \mathcal{F}\{⧢_T(t)\} \tag{23.56}$$

where

$$\mathcal{F}\{⧢_T(t)\} = \sum_{k=-\infty}^{\infty} \delta\left(\omega - \frac{2\pi}{T}k\right) = \frac{T}{2\pi} ⧢_{2\pi/T}(\omega) \tag{23.57}$$

so that

$$X_\Delta(i\omega) = \mathcal{F}\{x(t)\} * \mathcal{F}\{\sqcup\!\sqcup_T(t)\} = \sum_{k=-\infty}^{\infty} X\left[i\left(\omega - \frac{2\pi}{T}k\right)\right] \tag{23.58}$$

Thus, the Fourier transform X_Δ of the sampled variable has a periodic extension of the original spectrum $X(i\omega)$ along the frequency axis with a period equal to the sampling frequency $\omega_s = 2\pi/T$. There is an important result based on this observation known as the *Shannon sampling theorem*, which states that the continuous-time variable $x(t)$ may be reconstructed from the samples $\{x_k\}_{-\infty}^{+\infty}$ if and only if the sampling frequency is at least twice that of the highest frequency for which $X(i\omega)$ is nonzero. The original variable $x(t)$ may thus be recovered as

$$x(t) = \sum_{k=-\infty}^{\infty} x_k \frac{\sin\frac{\pi}{T}(t-kT)}{\frac{\pi}{T}(t-kT)} \tag{23.59}$$

The formula given in Eq. (23.59) is called *Shannon interpolation*, which is often quoted though it is valid only for infinitely long data sequences and would require a noncausal filter to reconstruct the continuous-time signal $x(t)$ in real-time operation. The frequency $\omega_n = \omega_s/2 = \pi/T$ is called the *Nyquist frequency* and indicates the upper limit of distortion-free sampling. A nonzero spectrum beyond this limit leads to interference between the sampling frequency and the sampled signal (*aliasing*); see Fig. 23.17.

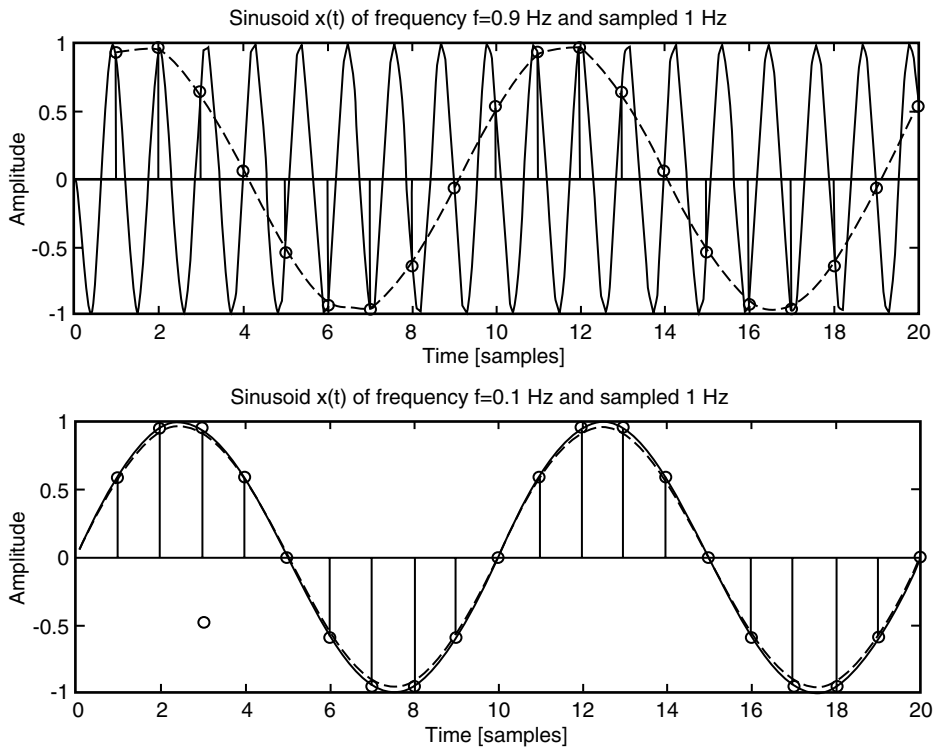



FIGURE 23.17 Illustration of aliasing appearing during sampling of a sinusoid $x(t) = \sin 2\pi \cdot 0.9t$ at the insufficient sampling frequency 1 Hz (sampling period $T = 1$) (*upper graph*). The sampled signal exhibits aliasing with its major component similar to a signal $x(t) = \sin 2\pi \cdot 0.1t$ sampled with the same rate (*lower graph*).

The Discrete Fourier Transform

Consider a finite length sequence $\{x_k\}_{k=0}^{N-1}$ that is zero outside the interval $0 \le k \le N-1$. Evaluation of the z transform $X(z)$ at N equally spaced points on the unit circle $z = \exp(i\omega_k T) = \exp[i(2\pi/NT)kT]$ for $k = 0, 1,\ldots, N-1$ defines the *discrete Fourier transform* (DFT) of a signal x with a sampling period h and N measurements:

$$X_k = \mathrm{DFT}\{x(kT)\} = \sum_{l=0}^{N-1} x_l \exp(-i\omega_k lT) = X(e^{i\omega_k T}) \tag{23.60}$$

Notice that the discrete Fourier transform $\{X_k\}_{k=0}^{N-1}$ is only defined at the discrete frequency points

$$\omega_k = \frac{2\pi}{NT}k, \quad \text{for } k = 0, 1,\ldots, N-1 \tag{23.61}$$

In fact, the discrete Fourier transform adapts the Fourier transform and the z transform to the practical requirements of finite measurements. Similar properties hold for the discrete Laplace transform with $z = \exp(sT)$, where s is the Laplace transform variable.

The Transfer Function

Consider the following discrete-time linear system with input sequence $\{u_k\}$ (stimulus) and output sequence $\{y_k\}$ (response). The dependency of the output of a linear system is characterized by the convolution-type equation and its z transform,

$$\begin{aligned} y_k &= \sum_{m=0}^{\infty} h_m u_{k-m} + v_k = \sum_{m=-\infty}^{k} h_{k-m} u_m + v_k, \quad k = \ldots, -1, 0, 1, 2,\ldots \\ Y(z) &= H(z)U(z) + V(z) \end{aligned} \tag{23.62}$$

where the sequence $\{v_k\}$ represents some external input of errors and disturbances and with $Y(z) = \mathcal{Z}\{y\}$, $U(z) = \mathcal{Z}\{u\}$, $V(z) = \mathcal{Z}\{v\}$ as output and inputs. The *weighting function* $h(kT) = \{h_k\}_{k=0}^{\infty}$, which is zero for negative k and for reasons of causality is sometimes called *pulse response* of the digital system (compare *impulse response* of continuous-time systems). The pulse response and its z transform, the *pulse transfer function*,

$$H(z) = \mathcal{Z}\{h(kT)\} = \sum_{k=0}^{\infty} h_k z^{-k} \tag{23.63}$$

determine the system's response to an input $U(z)$; see Fig. 23.18. The pulse transfer function $H(z)$ is obtained as the ratio

$$H(z) = \frac{X(z)}{U(z)} \tag{23.64}$$

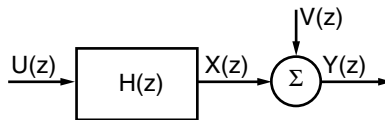



FIGURE 23.18 Block diagram with an assumed transfer function relationship $H(z)$ between input $U(z)$, disturbance $V(z)$, intermediate $X(z)$, and output $Y(z)$.

and provides the frequency domain input–output relation of the system. In particular, the Bode diagram is evaluated as $|H(z)|$ and arg $H(z)$ for $z = \exp(i\omega_k T)$ and for $|\omega_k| < \omega_n = \pi/T$, that is, when $H(z)$ is evaluated for frequency points up to the Nyquist frequency ω_n along the unit circle.

State-Space Systems

Alternatives to the input–output representations by means of transfer functions are the state-space representations. Consider the following finite dimensional discrete state-space equation with a state vector $x_k \in \mathbb{R}^n$, input $u_k \in \mathbb{R}^p$, and observations $y_n \in \mathbb{R}^m$.

$$\begin{cases} x_{k+1} = \Phi x_k + \Gamma u_k \\ & k = 0, 1, \ldots \\ y_k = C x_k + D u_k \end{cases} \tag{23.65}$$

with the pulse transfer function

$$H(z) = C(zI - \Phi)^{-1}\Gamma + D \tag{23.66}$$

and the output variable

$$Y(z) = C\sum_{k=0}^{\infty}\Phi^k z^{-k} x_0 + H(z)U(z) \tag{23.67}$$

where possible effects of initial conditions x_0 appear as the first term. Notice that the initial conditions x_0 can be viewed as the net effects of the input in the time interval $(-\infty, 0)$.

Digital Systems Described by Difference Equations (ARMAX Models)

An important class of nonstationary stochastic processes is one in which some deterministic response to an external input and a stationary stochastic process are superimposed. This is relevant, for instance, when the external input cannot be effectively described by some probabilistic distribution. A discrete-time model can be formulated in the form of a difference equation with an external input $\{u_k\}$ that is usually considered to be known:

$$y_k = -a_1 y_{k-1} - \cdots - a_n y_{k-n} + b_1 u_{k-1} + \cdots + b_n u_{k-n} + w_k + c_1 w_{k-1} + \cdots + c_n w_{k-n} \tag{23.68}$$

Application of the z transform permits formulation of Eq. (23.68) as

$$A(z^{-1})Y(z) = B(z^{-1})U(z) + C(z^{-1})W(z) \tag{23.69}$$

where

$$\begin{aligned} A(z^{-1}) &= 1 + a_1 z^{-1} + \cdots + a_n z^{-n} \\ B(z^{-1}) &= 1 + b_1 z^{-1} + \cdots + b_n z^{-n} \\ C(z^{-1}) &= 1 + c_1 z^{-1} + \cdots + c_n z^{-n} \end{aligned} \tag{23.70}$$

Stochastic models including the A polynomial, according to Eqs. (23.69) and (23.70), are known as **autoregressive (AR) models** and models including the C polynomial are known as **moving-average (MA) models**, whereas the B polynomial determines the effects of the external input (X). Notice that the term *moving average* is here somewhat misleading, as there is no restriction that the coefficients should add to 1 or that the coefficients are nonnegative. An alternative description is *finite impulse response* or *all-zero filter.*

Thus, the full model of Eq. (23.69) is an **autoregressive moving average model** with external input (ARMAX) and its pulse transfer function $H(z) = B(z^{-1})/A(z^{-1})$ is stable if and only if the *poles*—that is, the complex numbers $z_1,\dots, z_n$ solving the equation $A(z^{-1}) = 0$—are strictly inside the unit circle, that is, $|z_i| < 1$. The *zeros* of the system—that is, the complex numbers $z_1,\dots, z_n$ solving the equation $B(z^{-1}) = 0$—may take on any value without any instability arising, although it is preferable to obtain zeros located strictly inside the unit circle, that is, $|z_i| < 1$ (*minimum-phase zeros*). By linearity $\{y_k\}$ can be separated into one purely deterministic process $\{x_k\}$ and one purely stochastic process $\{v_k\}$:

$$\begin{cases} A(z^{-1})X(z) = B(z^{-1})U(z) \\ A(z^{-1})V(z) = C(z^{-1})W(z) \end{cases} \quad \text{and} \quad \begin{cases} y_k = x_k + v_k \\ Y(z) = X(z) + V(z) \end{cases} \tag{23.71}$$

The type of decomposition (Eq. (23.71)) that separates the deterministic and stochastic processes is known as the *Wold decomposition.*

Prediction and Reconstruction

Consider the problem of predicting the output d steps ahead when the output $\{y_k\}$ is generated by the ARMA model,

$$A(z^{-1})Y(z) = C(z^{-1})W(z) \tag{23.72}$$

which is driven by a zero-mean white noise $\{w_k\}$ with covariance $\mathscr{E}\{w_i w_j\} = \sigma_w^2 \delta_{ij}$. In other words, assuming that observations $\{y_k\}$ are available up to the present time, how should the output d steps ahead be predicted optimally? Assume that the polynomials $A(z^{-1})$ and $C(z^{-1})$ are mutually prime with no zeros for $|z| \geq 1$. Let the C polynomial be expanded according to the *Diophantine equation*,

$$C(z^{-1}) = A(z^{-1})F(z^{-1}) + z^{-d}G(z^{-1}) \tag{23.73}$$

which is solved by the two polynomials

$$\begin{aligned} F(z^{-1}) &= 1 + f_1 z^{-1} + \cdots + f_{nF} z^{-nF}, \quad & n_F &= d - 1 \\ G(z^{-1}) &= g_0 + g_1 z^{-1} + \cdots + g_{nG} z^{-nG}, \quad & n_G &= \max(n_A - 1, n_C - d) \end{aligned} \tag{23.74}$$

Interpretation of z^{-1} as a *backward shift operator* and application of Eqs. (23.72) and (23.73) permit the formulation

$$y_{k+d} = F(z^{-1})w_{k+d} + \frac{G(z^{-1})}{C(z^{-1})} y_k \tag{23.75}$$

Let us, by $\hat{y}_{k+d|k}$, denote linear d-step predictors of y_{k+d} based upon the measured information available at time k. As the zero-mean term $F(z^{-1})w_{k+d}$ of Eq. (23.75) is unpredictable at time k, it is natural to suggest the following d-step predictor:

$$\hat{y}_{k+d|k} = \frac{G(z^{-1})}{C(z^{-1})} y_k \tag{23.76}$$

The prediction error satisfies

$$\begin{aligned}
\varepsilon_{k+d} &= (\hat{y}_{k+d|k} - y_{k+d}) \\
&= \frac{G(z^{-1})}{C(z^{-1})} y_k - \frac{A(z^{-1})F(z^{-1}) + z^{-d}G(z^{-1})}{C(z^{-1})} y_{k+d} \\
&= -F(z^{-1})w_{k+d}
\end{aligned} \tag{23.77}$$

Let $\mathscr{E}\{\cdot|\mathscr{F}_k\}$ denote the *conditional mathematical expectation* relative to the measured information available at time k. The conditional mathematical expectation and the covariance of the d-step prediction relative to available information at time k is

$$\begin{aligned}
\mathscr{E}\{\hat{y}_{k+d|k} - y_{k+d} | \mathscr{F}_k\} &= \mathscr{E}\{-F(z^{-1})w_{k+d} | \mathscr{F}_k\} = 0 \\
\mathscr{E}\{(\hat{y}_{k+d|k} - y_{k+d})^2 | \mathscr{F}_k\} &= \mathscr{E}\{[F(z^{-1})w_{k+d}]^2 | \mathscr{F}_k\} \\
&= \mathscr{E}\{(w_{k+d} + f_1 w_{k+d-1} + \cdots + f_{d-1} w_{k+1})^2 | \mathscr{F}_k\} \\
&= (1 + f_1^2 + \cdots + f_{nF}^2)\sigma_w^2 = 0
\end{aligned} \tag{23.78}$$

It follows that the predictor of Eq. (23.76) is unbiased and that the prediction error only depends on future, unpredictable noise components. It is straightforward to show that the predictor of Eq. (23.76) achieves the lower bound of Eq. (23.78) and that the predictor of Eq. (23.76) is optimal in the sense that the prediction error variance is minimized.

Example 23.1—An Optimal Predictor for a First-Order Model

Consider for the first-order ARMA model

$$y_{k+1} = -a_1 y_k + w_{k+1} + c_1 w_k \tag{23.79}$$

The variance of a one-step-ahead predictor $\hat{y}_{k+1|k}$ is

$$\begin{aligned}
\mathscr{E}\{(\hat{y}_{k+1|k} - y_{k+1})^2 | \mathscr{F}_k\} &= \mathscr{E}\{(\hat{y}_{k+1|k} + a_1 y_k - c_1 w_k)^2 | \mathscr{F}_k\} + \mathscr{E}\{w_{k+1}^2 | \mathscr{F}_k\} \\
&= \mathscr{E}\{(\hat{y}_{k+1|k} + a_1 y_k - c_1 w_k)^2 | \mathscr{F}_k\} + \sigma_w^2 \geq \sigma_w^2
\end{aligned} \tag{23.80}$$

The optimal predictor satisfying the lower bound in Eq. (23.80) is obtained from Eq. (23.80) as

$$\hat{y}^o_{k+1|k} = -a_1 y_k + c_1 w_k \tag{23.81}$$

which, unfortunately, is not realizable as it stands because w_k is not available to measurement. Therefore, the noise sequence $\{w_k\}$ has to be substituted by some function of the observed variable $\{y_k\}$. A linear predictor chosen according to Eq. (23.76) is

$$\hat{y}_{k+1|k} = \frac{G(z^{-1})}{C(z^{-1})} y_k = \frac{c_1 - a_1}{1 + c_1 z^{-1}} y_k \tag{23.82}$$

The Kalman Filter

Consider the linear state-space model

$$\begin{aligned} x_{k+1} &= \Phi x_k + v_k, \quad x_k \in \mathbb{R}^n \\ y_k &= C x_k + w_k, \quad y_k \in \mathbb{R}^m \end{aligned} \tag{23.83}$$

where $\{v_k\}$ and $\{w_k\}$ are assumed to be independent zero-mean white-noise processes with covariances Σ_v and Σ_w, respectively. It is assumed that $\{y_k\}$, but not $\{x_k\}$, is available to measurement and that it is desirable to predict $\{x_k\}$ from measurements of $\{y_k\}$.

Introduce the state predictor,

$$\begin{aligned} \hat{x}_{k+1|k} &= \Phi \hat{x}_{k|k-1} - K_k(\hat{y}_k - y_k), \quad \hat{x}_{k|k-1} \in \mathbb{R}^n \\ \hat{y}_k &= C \hat{x}_{k|k-1}, \qquad y_k \in \mathbb{R}^m \end{aligned} \tag{23.84}$$

The predictor of Eq. (23.84) has the same dynamics matrix Φ as the state-space model of Eq. (101.83) and, in addition, there is a correction term $K_k(\hat{y}_k - y_k)$ with a factor K_k to be chosen. The prediction error is

$$\tilde{x}_{k+1|k} = \hat{x}_{k+1|k} - x_{k+1} \tag{23.85}$$

The prediction-error dynamics is

$$\tilde{x}_{k+1} = (\Phi - K_k C)\tilde{x}_k + v_k - K_k w_k \tag{23.86}$$

The mean prediction error is governed by the recursive equation

$$\mathcal{E}\{\tilde{x}_{k+1}\} = (\Phi - K_k C)\mathcal{E}\{\tilde{x}_k\} \tag{23.87}$$

and the mean square error of the prediction error is governed by

$$\begin{aligned} \mathcal{E}\{\tilde{x}_{k+1}\tilde{x}_{k+1}^T\} &= \mathcal{E}\{[(\Phi - K_k C)\tilde{x}_k + v_k - K_k w_k][(\Phi - K_k C)\tilde{x}_k + v_k - K_k w_k]^T\} \\ &= (\Phi - K_k C)\mathcal{E}\{\tilde{x}_k \tilde{x}_k^T\}(\Phi - K_k C)^T + \Sigma_v + K_k \Sigma_w K_k \end{aligned} \tag{23.88}$$

If we denote

$$P_k = \mathcal{E}\{\tilde{x}_k \tilde{x}_k^T\}, \qquad Q_k = \Sigma_w + C P_k C^T \tag{23.89}$$

then Eq. (23.88) is simplified to

$$P_{k+1} = \Phi P_k \Phi^T - K_k C P_k \Phi - \Phi^T P_k C^T K_k^T + \Sigma_v + K_k Q_k K_k^T \tag{23.90}$$

By completing squares of terms containing K_k we find

$$P_{k+1} = \Phi P_k \Phi^T + \Sigma_v - \Phi P_k C^T Q_k^{-1} C P_k \Phi^T + (K_k - \Phi P_k C^T Q_k^{-1}) Q_k (K_k - \Phi P_k C^T Q_k^{-1})^T \qquad (23.91)$$

where only the last term depends on K_k. Minimization of P_{k+1} can be done by choosing K_k such that the positive semidefinite K_k-dependent term in Eq. (23.91) disappears. Thus P_{k+1} achieves its lower bound for

$$K_k = \Phi P_k C^T (\Sigma_w + C P_k C^T)^{-1} \qquad (23.92)$$

and the *Kalman filter* (or *Kalman–Bucy filter*) takes the form

$$\begin{aligned} \hat{x}_{k+1|k} &= \Phi \hat{x}_{k|k-1} - K_k(\hat{y}_k - y_k) \\ \hat{y}_k &= C\hat{x}_{k|k-1}, \qquad K_k = \Phi P_k C^T (\Sigma_w + C P_k C^T)^{-1} \\ P_{k+1} &= \Phi P_k \Phi^T + \Sigma_v - \Phi P_k C^T (\Sigma_w + C P_k C^T)^{-1} C P_k \Phi^T \end{aligned} \qquad (23.93)$$

which is the optimal predictor in the sense that the mean square error (Eq. (23.88)) is minimized in each step.

Example 23.2—Kalman Filter for a First-Order System

Consider the state-space model

$$x_{k+1} = 0.95 x_k + v_k, \qquad y_k = x_k + w_k \qquad (23.94)$$

where $\{v_k\}$ and $\{w_k\}$ are zero-mean white-noise processes with covariances $\mathcal{E}\{v_k^2 = 1$ and $\mathcal{E}\{w_k^2\} = 1$, respectively.

The Kalman filter takes on the form

$$\begin{aligned} \hat{x}_{k+1|k} &= 0.95\hat{x}_{k|k-1} - K_k(\hat{x}_{k|k-1} - y_k) \\ K_k &= \frac{0.95 P_k}{1 + P_k} \\ P_{k+1} &= 0.95^2 P_k + 1 - \frac{0.95^2 P_k^2}{1 + P_k} \end{aligned} \qquad (23.95)$$

The result of one such realization is shown in Fig. 23.19.

Defining Terms

Autoregressive (AR) model: An autoregressive time series of order n is defined via $y_k = -\Sigma_{m=1}^{n} a_m y_{k-m} + w_k$. The sequence $\{w_k\}$ is usually assumed to consist of zero-mean identically distributed stochastic variables w_k.

Autoregressive moving average (ARMA) model: An autoregressive moving average time series of order n is defined via $y_k = -\Sigma_{m=1}^{n} a_m y_{k-m} + \Sigma_{m=0}^{n} c_m w_{k-m}$. The sequence $\{w_k\}$ is usually assumed to consist of zero-mean identically distributed stochastic variables w_k.

Discrete Laplace transform: The discrete Laplace transform is a counterpart to the Laplace transform with application to discrete signals and systems. The discrete Laplace transform is obtained from the z transform by means of the substitution $z = \exp(sT)$, where T is the sampling period.

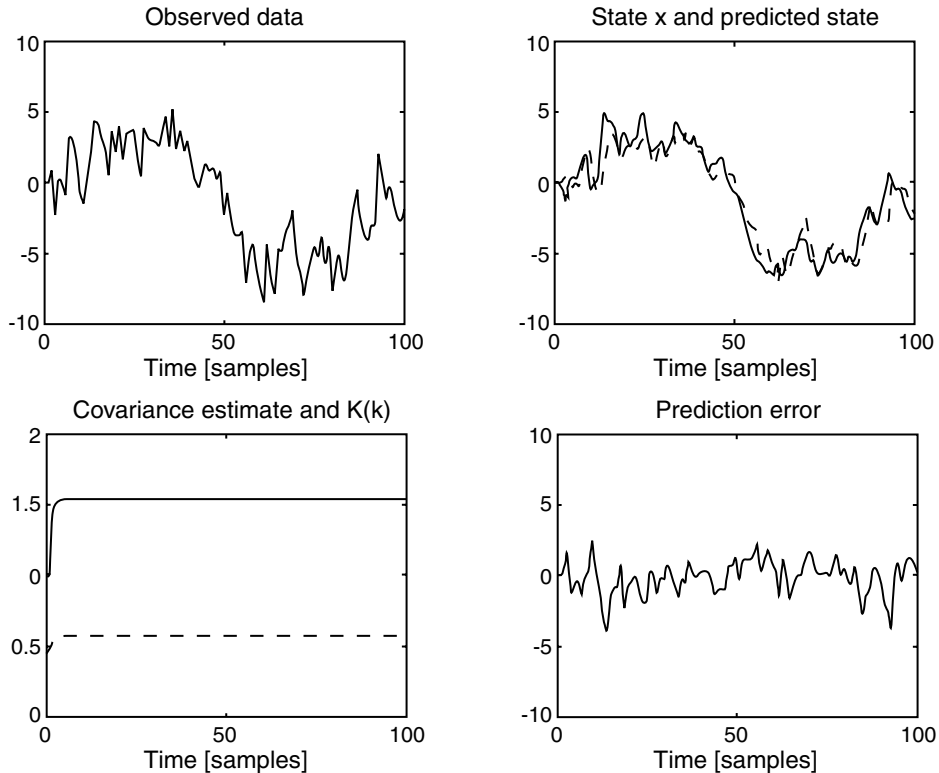



FIGURE 23.19 Kalman filter applied to one-step-ahead prediction of x_{k+1} in Eq. (23.94). The observed variable $\{y_k\}$, the state $\{x_k\}$, and the predicted state $\{\hat{x}_k\}$, the estimated variance $\{P_k\}$ and $\{K_k\}$, and the prediction error $\{\tilde{x}_k\}$ are shown in a 100-step realization of the stochastic process. (*Source*: Johansson, R. 1993. *System Modeling and Identification*. Prentice-Hall, Englewood Cliffs, NJ.)

Moving average (MA) process: A moving average time series of order n is defined via $y_k = \Sigma_{m=0}^{n} c_m w_{k-m}$. The sequence $\{w_k\}$ is usually assumed to consist of zero-mean identically distributed stochastic variables w_k.

Rational model: AR, MA, ARMA, and ARMAX are commonly referred to as rational models.

Time Series: A sequence of random variable $\{y_k\}$, where k belongs to the set of positive and negative integers.

***z* transform:** A generating function applied to sequences of data and evaluated as a function of the complex variable z with interpretation of frequency.

References

Box, G. E. P. and Jenkins, G. M. 1970. *Time Series Analysis: Forecasting and Control.* Holden-Day, San Francisco, CA.

Hurewicz, W. 1947. Filters and servo systems with pulsed data. In *Theory of Servomechanisms,* H. M. James, N. B. Nichols, and R. S. Philips, eds., McGraw-Hill, New York.

Jenkins, G. M. and Watts, D. G. 1968. *Spectral Analysis and Its Applications.* Holden-Day, San Francisco, CA.

Johansson, R. 1993. *System Modeling and Identification.* Prentice-Hall, Englewood Cliffs, NJ.

Jury, E. I. 1956. Synthesis and critical study of sampled-data control systems. *AIEE Trans.* 75: 141–151.

Kalman, R. E. and Bertram, J. E. 1958. General synthesis procedure for computer control of single and multi-loop linear systems. *Trans. AIEE.* 77: 602–609.

Kolmogorov, A. N. 1939. Sur l'interpolation et extrapolation des suites stationnaires. *C. R. Acad. Sci.* 208: 2043–2045.

Ragazzini, J. R. and Zadeh, L. A. 1952. The analysis of sampled-data systems. *AIEE Trans.* 71:225–234.

Tsypkin, Y. Z. 1950. Theory of discontinuous control. *Avtomatika i Telemekhanika.* Vol. 5.

Wiener, N. 1949. *Extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications.* John Wiley & Sons, New York.

Further Information

Early theoretical efforts developed in connection with servomechanisms and radar applications [Hurewicz, 1947]. Tsypkin [1950] introduced the discrete Laplace transform and the formal z transform definition was introduced by Ragazzini and Zadeh [1952] with further developments by Jury [1956]. Much of prediction theory was originally developed by Kolmogorov [1939] and Wiener [1949] whereas state-space methods were forwarded by Kalman and Bertram [1958]. Pioneering textbooks on time-series analysis and spectrum analysis are provided by Box and Jenkins [1970] and Jenkins and Watts [1968].

Detailed accounts of time-series analysis and the z transform and their application to signal processing are to be found in

- Oppenheim, A. V. and Schafer, R. W. 1989. *Discrete-Time Signal Processing.* Prentice-Hall, Englewood Cliffs, NJ.
- Proakis, J. G. and Manolakis, D. G. 1989. *Introduction to Digital Signal Processing.* Maxwell MacMillan Int. Ed., New York.

Theory of time-series analysis and its application to discrete-time control is to be found in

- Åström, K. J. and Wittenmark, B. 1990. *Computer-Controlled Systems,* 2nd ed., Prentice-Hall, Englewood Cliffs, NJ.

Theory of time-series analysis and methodology for determination and validation of discrete-time models and other aspects of system identification are to be found in

- Johansson, R. 1993. *System Modeling and Identification.* Prentice-Hall, Englewood Cliffs, NJ.

Good sources to monitor current research are

- *IEEE Transactions on Automatic Control*
- *IEEE Transactions on Signal Processing*

Examples of easy-to-read survey articles for signal processing applications are

- Cadzow, J. A. 1990. Signal processing via least-squares error modeling. *IEEE ASSP Magazine.* 7:12–31, October.
- Schroeder, M. R. 1984. Linear prediction, entropy, and signal analysis. *IEEE ASSP Magazine.* 1:3–11, July.

23.3 Continuous- and Discrete-Time State-Space Models

Kam Leang, Qingze Zou, and Santosh Devasia

Introduction

In this section we introduce the modeling of continuous- and discrete-time systems using the state-space approach. The state-space approach is a technique that uses a set of first order differential equations to represent the behavior of a system in the time-domain. The state-space approach has an advantage over frequency-domain approaches such as the transfer-function approach: it can be used to model linear,

nonlinear, time-varying, and multivariable systems, whereas the transfer-function approach is suited to linear time-invariant (LTI) systems [1, Chapter 3]. In addition, models expressed in first order state-space form in the time-domain can be readily solved by a digital computer or microprocessor, which makes this approach quite useful for the design and control of modern mechatronic systems. Furthermore, there is a wide variety of available computer software, such as MATLAB [2], that take advantage of the state-space form for analyzing and solving design problems. Therefore, the state-space approach can be used to investigate the behavior and facilitate in the design of both continuous- and discrete-time systems, the fundamentals of which will be the focus of this section.

In the following, we begin with an example: the modeling of a piezoceramic actuator and use the example throughout the section. The concept of a system state is introduced and we explain the state-space equation for linear systems and present its solution. The topic of linearization of nonlinear systems is briefly mentioned. The relationships between time- and frequency-domain models are discussed and a procedure for obtaining a state-space model using experimental frequency-domain (frequency-response) data is presented. This section closes with a discussion of discrete-time state-space modeling and concluding remarks. Useful MATLAB commands are also included as footnotes.

States and the State-Space

An Example Piezoceramic Actuator

We begin by modeling a piezoceramic actuator, which is an example mechatronic (electromechanical) system. When a voltage is applied to a piezoceramic material, its dimension changes. This change in dimension can be used to precisely position an object or tool (such as a sensor), therefore making piezoceramics suitable actuators for a wide variety of applications. For example, due to their ability to achieve positioning with sub-nanometer level precision, piezoceramic actuators have become ideal for emerging nanotechnologies. In particular, a piezo-tube actuator is used in scanning probe microscopes (SPMs, see Fig. 23.20) to precisely position a probe tip for high-precision nanofabrication, surface modification, and the acquisition of images of atoms [3]. The probe tip can be positioned in the three coordinate axes (x, y, and z), with each motion controlled by an independent voltage source ($V_x(t)$, $V_y(t)$, and $V_z(t)$). Scanning of the probe is performed parallel to the sample surface along the x- and y-axis; the z-axis movement allows motion of the probe perpendicular to the sample surface. An accurate mathematical model of the dynamics of a piezo-tube actuator is required for the analysis and design of SPM systems. A designer can exploit the known information of the system from its model to improve or optimize a design for building faster and more reliable SPMs. For example, an approach that has been

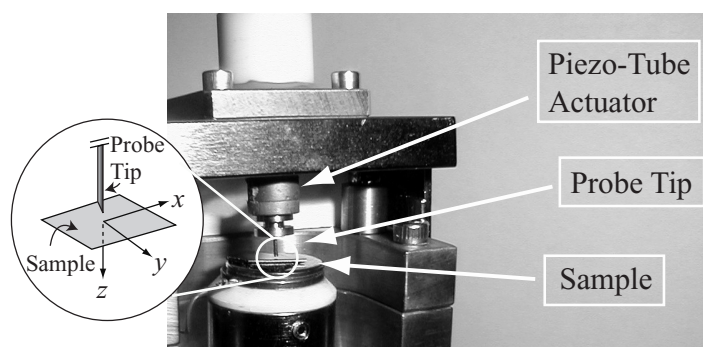



FIGURE 23.20 The main components of a scanning probe microscope (SPM) used for surface analysis, which includes the piezo-tube actuator, the probe tip, and the sample. The configuration of the probe tip and sample with respect to the coordinate axes (x, y, and z) are shown in the magnified view.

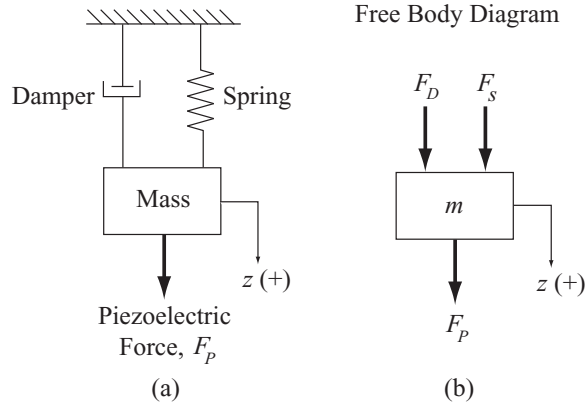



FIGURE 23.21 (a) A simple *lumped model* of the piezo-tube actuator modeled along the *z*-axis consisting of a mass, a spring, and a damper [4]. The positive *z*-direction is indicated by the arrow and the "+" sign. (b) The forces acting on the mass (free body diagram).

successfully implemented is the inversion-based control method, which finds the inputs required to achieve exact tracking by inverting the system model [3]. This technique works best when the dynamics of the system are well characterized and understood. In general, the analysis and design of control systems also requires a system model. Thus for analysis and design, it is crucial to obtain an accurate mathematical model that describes the behavior of a system. Modeling of the example piezo-tube system is considered in the following.

Simple Model of a Piezo-Tube Actuator

We will model the dynamics of the piezo-tube actuator along the *z*-axis where the input is the applied voltage $V_z(t)$ and the output of the system is the displacement of the probe tip $z(t)$. We begin the modeling by simplifying the system as an isolated mass, an ideal spring, and a damper as shown in Fig. 23.21(a). The entire mass of the piezo-tube is lumped into one mass element m, the internal elastic behavior of the piezo-tube is modeled as a spring, and the structural damping in the piezo-tube is modeled as a damper or a viscous friction element (such models are referred to as *lumped models* [4]). A mathematical relationship between the applied voltage $V_z(t)$ and the displacement of the probe tip $z(t)$ can be obtained using physical laws. Applying Newton's second law (the sum of all external forces F_i acting on a body is equal the product of its mass m and acceleration $\ddot{z}(t)$) we can write the equation of motion as

$$\sum_i F_i(t) = m\ddot{z}(t) \tag{23.96}$$

As shown in Fig. 23.21(b) (the free body diagram), there are three external forces acting on the piezo-tube. First, the force exerted by the spring is assumed to be proportional to the displacement of the probe tip, i.e.,

$$F_s(t) = -kz(t) \tag{23.97}$$

where k is the spring constant with SI units [N/m]. Second, the damping force is considered to be proportional to the velocity of the probe tip $\dot{z}(t)$, i.e.,

$$F_D(t) = -c\dot{z}(t) \tag{23.98}$$

where c is the viscous friction or damping coefficient with SI units [N · s/m]. Third, induced strain ε in the piezoceramic material is proportional to the applied voltage $V_z(t)$ [5], and by Hooke's Law, the induced stress σ is proportional to the induced strain ε. Hence, the induced force $F_p(t)$ (stress σ times the cross-sectional area) is proportional to the applied voltage $V_z(t)$, i.e.,

$$F_p(t) = bV_z(t) \tag{23.99}$$

where b is a constant with SI units [N/V]. Rewriting Eq. (23.96) in terms of the three external forces, the equation of motion becomes

$$\sum_{i=1}^{3} F_i(t) = F_S(t) + F_D(t) + F_p(t) = -kz(t) - c\dot{z}(t) + bV_z(t) = m\ddot{z}(t) \tag{23.100}$$

which is called the *mass-spring-damper* model. Note that the relationship between the input voltage $V_z(t)$ and the displacement $z(t)$ of the probe tip (i.e., the model of the dynamics) is a second order differential equation. The response of the probe tip (displacement of mass m) to an applied voltage $V_z(t)$ can be obtained in the frequency-domain by using the Laplace transform technique [6, Chapter 2, section 5]; however, the state-space approach can be used to obtain the solution directly in the time-domain. In the remaining sections, the state-space approach to modeling is presented and the *mass-spring-damper* model of the piezo-tube actuator will be used as an example.

States of a System

We begin by introducing the concept of a state, which is the basis for the state-space approach. In general, a state can be defined as the following:

> The state $x(t_0)$ of a dynamic system at time t_0 is a set of variables that, together with the input $u(t)$, for $t \geq t_0$, determines the behavior of the system for all $t \geq t_0$ [7, Chapter 2, section 1.1].

Fundamental to this definition is the notion that the state summarizes the current configuration of a system. Therefore, the memory of a dynamical system is preserved in the state variables at the current time t_0 (called initial condition), and the future behavior of the system is determined by the initial condition $x(t_0)$ and the applied input $u(t)$, for $t \geq t_0$. The state of a system can be written as the set

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_n(t) \end{bmatrix} \tag{23.101}$$

where n is the number of states.[1] Any set of variables that satisfy the above definition can be a valid state, hence the state is not unique [8, Chapter 2, section 2].

Example

The state variables required to describe the mass-spring-damper system can be chosen as the position $z(t)$ and velocity $\dot{z}(t)$ of the mass. We can write the state vector as

$$x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} z(t) \\ \dot{z}(t) \end{bmatrix} \tag{23.102}$$

[1]For a discussion on the minimal set of states required to describe a system (minimal realization), see [7, Chapter 7].

where the number of states is two ($n = 2$). If the position $z(t)$ and velocity $\dot{z}(t)$ of the mass are known at time t_0, along with the applied voltage $V_z(t)$ defined for $t \geq t_0$, then the future behavior of the system (i.e., the state $x(t)$) can be determined by solving the differential Eq. (23.100).

The Linear State-Space Equation and Its Solution

For a linear system, the evolution of the states of a system over time can be described by a set of linear first order differential equations of the form:

$$\begin{aligned} \dot{x}_1(t) &= \frac{dx_1(t)}{dt} = a_{11}(t)x_1(t) + \cdots + a_{1n}(t)x_n(t) + b_{11}(t)u_1(t) + \cdots b_{1p}(t)u_p(t) \\ \dot{x}_2(t) &= \frac{dx_2(t)}{dt} = a_{21}(t)x_1(t) + \cdots + a_{2n}(t)x_n(t) + b_{21}(t)u_1(t) + \cdots b_{2p}(t)u_p(t) \\ &\vdots \\ \dot{x}_n(t) &= \frac{dx_n(t)}{dt} = a_{n1}(t)x_1(t) + \cdots + a_{nn}(t)x_n(t) + b_{n1}(t)u_1(t) + \cdots b_{np}(t)u_p(t) \end{aligned} \tag{23.103}$$

where n is the number of states (or the order of the system) and p is the number of inputs.[2] Defining the input vector as

$$u(t) = \begin{bmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_p(t) \end{bmatrix} \tag{23.104}$$

and the state vector $x(t)$ as defined in Eq. (23.101), the set of first order differential equations given by Eq. (23.103) can be rewritten in compact matrix form as [8, Chapter 2, section 2]

$$\begin{aligned} \dot{x}(t) &= \begin{bmatrix} a_{11}(t) & a_{12}(t) & \cdots & a_{1n}(t) \\ a_{21}(t) & a_{22}(t) & \cdots & a_{2n}(t) \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1}(t) & a_{n2}(t) & \cdots & a_{nn}(t) \end{bmatrix} x(t) + \begin{bmatrix} b_{11}(t) & b_{12}(t) & \cdots & b_{1p}(t) \\ b_{21}(t) & b_{22}(t) & \cdots & b_{2p}(t) \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1}(t) & b_{n2}(t) & \cdots & b_{np}(t) \end{bmatrix} u(t) \\ &= A(t)x(t) + B(t)u(t) \end{aligned} \tag{23.105}$$

where $A(t)$ is an $n \times n$ matrix and $B(t)$ is an $n \times p$ matrix. For a system defined with q outputs $y(t)$, which are assumed to be a linear combination of the state $x(t)$ and input $u(t)$, we can write the output equation as

$$\begin{aligned} y(t) &= \begin{bmatrix} c_{11}(t) & c_{12}(t) & \cdots & c_{1n}(t) \\ c_{21}(t) & c_{22}(t) & \cdots & c_{2n}(t) \\ \vdots & \vdots & \ddots & \vdots \\ c_{q1}(t) & c_{q2}(t) & \cdots & c_{qn}(t) \end{bmatrix} x(t) + \begin{bmatrix} d_{11}(t) & d_{12}(t) & \cdots & d_{1p}(t) \\ d_{21}(t) & d_{22}(t) & \cdots & d_{2p}(t) \\ \vdots & \vdots & \ddots & \vdots \\ d_{q1}(t) & d_{q2}(t) & \cdots & d_{qp}(t) \end{bmatrix} u(t) \\ &= C(t)x(t) + D(t)u(t) \end{aligned} \tag{23.106}$$

[2]Given a higher order differential equation, a set of first order differential equations can be obtained by a procedure known as reduction to first order as presented in [9].

where $C(t)$ is a $q \times n$ matrix and $D(t)$ is a $q \times p$ matrix. In general, the matrices $A(t)$, $B(t)$, $C(t)$, and $D(t)$ are time varying; however, in this chapter we will only consider the time-invariant case where A, B, C, and D are constant matrices, then Eqs. (23.107) and (23.108) are called the linear time-invariant (LTI) state and output equations, respectively.[3]

$$\dot{x}(t) = Ax(t) + Bu(t) \tag{23.107}$$

$$y(t) = Cx(t) + Du(t) \tag{23.108}$$

The response of the system to an applied input can be quantified by the evolution of the system state $x(t)$ and the output $y(t)$. The state-space Eq. (23.107) is a set of first order differential equations in matrix form, which can be solved in time for a given initial condition $x(t_0)$ as [8, Chapter 3]

$$x(t) = e^{A(t-t_0)}x(t_0) + \int_{t_0}^{t} e^{A(t-\tau)}Bu(\tau)d\tau \tag{23.109}$$

Note that solution (23.109) is the sum of two terms: the first term is the effect of initial condition $x(t_0)$ and the second is the effect of the applied input $u(t)$ between $t_0 \le \tau \le t$.[4] Using the output Eq. (23.108) and the state solution given by (23.109), the output $y(t)$ becomes

$$y(t) = Ce^{A(t-t_0)}x(t_0) + \int_{t_0}^{t} Ce^{A(t-\tau)}Bu(\tau)d\tau + Du(t) \tag{23.110}$$

The system response $y(t)$ to an applied input $u(t)$ is characterized by the system matrices (A, B, C, D). For example, the output $y(t)$ will be bounded for any bounded input if the system is stable and the system is stable if the real parts of all the eigenvalues of A are less than zero (strictly negative) [8, Chapter 4, section 4].[5]

Example

For the mass-spring-damper example system, the state-space equation can be found by differentiating the states $x(t)$ defined in Eq. (23.102) and using the equation of motion (23.100) to obtain

$$\begin{aligned} \dot{x}_1(t) = \dot{z}(t) &= x_2(t) \\ \dot{x}_1(t) = \ddot{z}(t) &= -\left(\frac{k}{m}\right)z(t) - \left(\frac{c}{m}\right)\dot{z}(t) + \left(\frac{b}{m}\right)V_z(t) = -\left(\frac{k}{m}\right)x_1(t) - \left(\frac{c}{m}\right)x_2(t) + \left(\frac{b}{m}\right)u(t) \end{aligned} \tag{23.111}$$

We choose the position of the mass $z(t)$ to be the output of the system, and write the state-space and output equation in the form given by Eqs. (23.107) and (23.108) as

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ -(k/m) & -(c/m) \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ b/m \end{bmatrix} u(t) \tag{23.112}$$

$$y(t) = \begin{bmatrix} 1 & 0 \end{bmatrix} x(t) \tag{23.113}$$

[3]For a detailed discussion of the solution of linear time-varying equations, see [7, Chapter 4, section 5].
[4]The MATLAB command `lsim` simulates the time response of LTI models to arbitrary inputs.
[5]The MATLAB command `eig (A)` returns the eigenvalues of the system matrix `A`.

Linearization of Nonlinear Systems

A general form of the state-space equation (for nonlinear systems) is

$$\dot{x}(t) = g(x, u) \tag{23.114}$$

$$y(t) = h(x, u) \tag{23.115}$$

where g and h can be nonlinear functions.[6] The behavior of nonlinear systems is beyond the scope of this section; however, a detailed discussion can be found in [10]. The behavior of a nonlinear system can be approximated by a linear model in a neighborhood of an equilibrium point. Such linearizations can simplify the analysis and design of nonlinear systems because the tools developed for linear systems can be applied under certain conditions [10]. Let x_0 and u_0 be the equilibrium point and equilibrium input, respectively, such that [10, Chapter 1]

$$g(x_0, u_0) = 0 \tag{23.116}$$

$$h(x_0, u_0) = y_0 \tag{23.117}$$

Consider small perturbations in the equilibrium point $x(t) = x_0 + \bar{x}(t)$, the input $u(t) = u_0(t) + \bar{u}(t)$, and the output $y(t) = y_0 + \bar{y}(t)$. If the perturbation $\bar{x}(t)$ is small for all t, we obtain the following by expanding (23.114) in Taylor series (neglecting higher order terms of $\bar{x}(t)$ and $\bar{u}(t)$):

$$\begin{aligned} \dot{x}_0 + \dot{\bar{x}}(t) &= g(x_0 + \bar{x}(t), u_0 + \bar{u}(t)) \\ \dot{\bar{x}}(t) &= g(x_0, u_0) + \left.\frac{\partial g}{\partial x}\right|_{\substack{x=x_0 \\ u=u_0}} \bar{x}(t) + \left.\frac{\partial g}{\partial u}\right|_{\substack{x=x_0 \\ u=u_0}} \bar{u}(t) \end{aligned} \tag{23.118}$$

Recognizing that $g(x_0, x_0) = 0$ we obtain

$$\dot{\bar{x}}(t) = \bar{A}\bar{x}(t) + \bar{B}\bar{u}(t) \tag{23.119}$$

where

$$\bar{A} = \left.\frac{\partial g}{\partial x}\right|_{\substack{x=x_0 \\ u=u_0}} \quad \text{and} \quad \bar{B} = \left.\frac{\partial g}{\partial u}\right|_{\substack{x=x_0 \\ u=u_0}} \tag{23.120}$$

The matrices $\bar{A}$ and $\bar{B}$ are the Jacobians evaluated at x_0 and u_0. Equation (23.119) is a linear state equation and is valid for small perturbations about x_0 and u_0. A similar result can be obtained for the change $\bar{y}(t)$ in the output from the equilibrium value y_0 as

$$\bar{y}(t) = \bar{C}\bar{x}(t) + \bar{D}\bar{u}(t) \tag{23.121}$$

where

$$\bar{C} = \left.\frac{\partial h}{\partial x}\right|_{\substack{x=x_0 \\ u=u_0}} \quad \text{and} \quad \bar{D} = \left.\frac{\partial h}{\partial u}\right|_{\substack{x=x_0 \\ u=u_0}} \tag{23.122}$$

[6]The MATLAB command `ode45` can be used to obtain the numeric solution to the general nonlinear state space equation.

Relationship between State Equations and Transfer-Functions

State-Space to Transfer-Function

The input-to-output relationship of a dynamic system in the frequency-domain is represented by a transfer-function, which can be obtained by taking the Laplace transform of (23.107) and (23.108) with zero initial conditions as follows [8, Chapter 3, section 5]:

$$sX(s) = AX(s) + BU(s), \tag{23.123}$$

$$Y(s) = CX(s) + DU(s), \tag{23.124}$$

where s is the Laplace variable. Solving (23.123) for $X(s)$ and substituting into (23.124), the ratio of the output $Y(s)$ to input $U(s)$ for a single-input single-output system (SISO) can be found as

$$\begin{aligned} G(s) &= \frac{Y(s)}{U(s)} = C(sI - A)^{-1}B + D \\ &= \frac{N(s)}{D(s)} \end{aligned} \tag{23.125}$$

where I is an $n \times n$ identity matrix. In Eq. (23.125), $N(s)$ and $D(s)$ are referred to as the numerator and denominator polynomial of $G(s)$, respectively.[7]

Analogous to the state-space equation, the boundedness of the output response $y(t)$ to a bounded input $u(t)$ is characterized by the roots of the denominator polynomial $D(s)$, i.e., the values of s for which $D(s) = 0$. In particular, the output $y(t)$ will be bounded for any bounded input, i.e., system is stable if the real parts of all the roots of $D(s)$ are less than zero (strictly negative).[8] Alternatively, a convenient method to determine stability without having to find the roots of $D(s)$ explicitly is the Routh–Hurwitz stability criterion [6, Chapter 6].

Example

With the state-space description of the mass-spring-damper system defined in Eqs. (23.112) and (23.113), the transfer-function realization using Eq. (23.125) becomes

$$\begin{aligned} G(s) &= \frac{Y(s)}{U(s)} = \begin{bmatrix} 1 & 0 \end{bmatrix} \left[s\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 1 \\ -(k/m) & -(c/m) \end{bmatrix} \right]^{-1} \begin{bmatrix} 0 \\ b/m \end{bmatrix} + [0] \\ &= \frac{b/m}{s^2 + (c/m)s + k/m} \end{aligned} \tag{23.126}$$

The input to the system is the applied voltage $V_z(t)$ and the output is the displacement of the mass $z(t)$.

Frequency-Response Using Transfer-Functions

Consider a linear single-input single-output (SISO) stable system with transfer-function description $G(s)$. When the system $G(s)$ is excited by a sinusoidal input of the form

$$u(t) = P\sin(\omega t) \tag{23.127}$$

with amplitude P and frequency ω, the output response (after the transients decay) will also be a sinusoid of the form

$$y(t) = MP\sin(\omega t + \phi) \tag{23.128}$$

[7]The MATLAB command `ss2tf` can be used to convert a state-space realization to a transfer-function.

[8]The MATLAB command `roots (den)` can be used to find the roots of `den`, where `den` is the coefficients of $D(s)$.

with the same frequency ω and a phase shift ϕ [6, Chapter 8]. The output amplitude is the input amplitude scaled by M, the magnitude gain. The magnitude gain M is found by taking the magnitude of $G(s)$ evaluated at $s = j\omega$, i.e.,

$$M = |G(s)|_{s=j\omega} \tag{23.129}$$

Usually, the magnitude gain M is expressed in units of decibels (dB), i.e., M [dB] = 20log M. The phase shift ϕ is the angle of $G(s)$ evaluated at $s = j\omega$, i.e.,

$$\phi = \angle G(s)|_{s=j\omega} \tag{23.130}$$

with units of degrees. The plot of the magnitude gain M and the phase shift ϕ versus the frequency ω gives a graphical representation of the frequency-response (Bode plots) of a system.[9] These plots can be obtained experimentally by measuring the magnitude gain and phase shift between the input and output response of a system over a range of frequencies. Additionally, a system's transfer-function can be obtained from an experimental frequency-response data by using curve-fitting software. In section "Experimental Modeling Using Frequency Response," we present this approach to determine a model for a system using experimental frequency-response data.

Transfer-Function to State-Space

In section "State-Space to Transfer-Function," a transfer-function model was obtained for a system in state-space form. In the following, an approach for realizing a state-space model from a transfer-function $G(s)$ is presented. For a realizable transfer function $G(s)$ of a SISO system of the form

$$G(s) = \frac{b_0 s^n + b_1 s^{n-1} + \cdots + b_n}{s^n + a_1 s^{n-1} + \cdots + a_n} \tag{23.131}$$

the controllable canonical state-space form written in terms of the coefficients of $G(s)$ is

$$\dot{x}(t) = \begin{bmatrix} -a_1 & -a_2 & \cdots & -a_{n-1} & -a_n \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} u(t) \tag{23.132}$$

$$y(t) = [(b_1 - a_1 b_0) \;\; (b_2 - a_2 b_0) \;\; \cdots \;\; (b_n - a_n b_0)] x(t) + [b_0] u(t) \tag{23.133}$$

where the number of states n is equal to the highest power of the denominator of $G(s)$.[10] The smallest possible dimension for realizing a system, referred to as the minimum realization, is an important factor to consider in analysis and design.[11] Models of minimum order require less computational power in simulation and implementation compared to higher order models. For information about other equivalent canonical state-space forms, refer to [7, Chapter 4, sections 3 and 4].

---

[9] The MATLAB command `bode` plots the magnitude gain and phase shift for a linear system.

[10] The MATLAB command `tf2ss` generates the controllable canonical form realization of a transfer function $G(s)$. Other useful commands include `ss2tf`, `zp2tf`, and `tf2zp`.

[11] For a detailed discussion of minimal realizations for multi-input multi-output systems, see [7, Chapter 7].

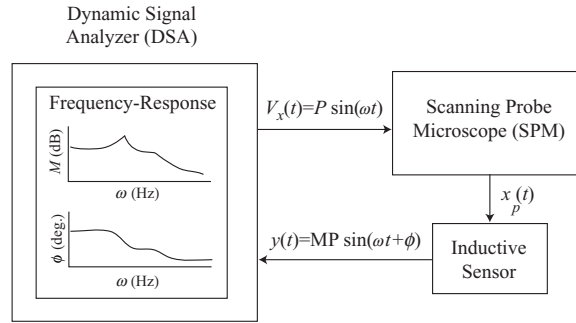



FIGURE 23.22 A schematic of the experimental setup used to determine the frequency-response of the piezo-tube actuator. An inductive sensor measured the displacement of the actuator along the x-axis, and the frequency-response data from the DSA were used to estimate the system model.

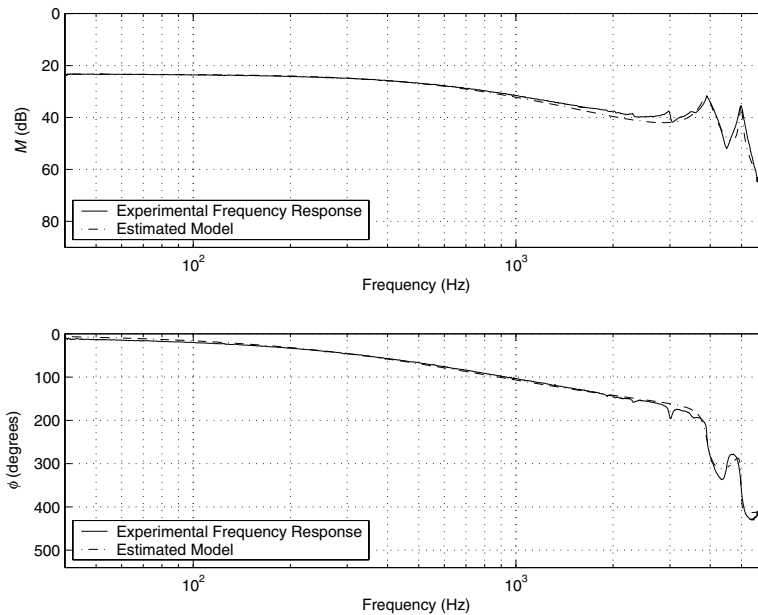



FIGURE 23.23 The experimental magnitude gain and phase versus frequency plots for the piezo-tube actuator measured along the x-axis with superimposed model frequency-response. Solid line represents experimental data; dashed line represents results from estimated model.

Experimental Modeling Using Frequency-Response

An approach to modeling using experimental frequency-response data is presented in this section. Using a dynamic signal analyzer (DSA), the frequency-response of the dynamics along the x-axis for the piezo-tube actuator was measured.[12] A sinusoidal input voltage $V_x(t)$ with frequency varying between 10 Hz and 6 kHz was generated by a DSA and applied to the scanning probe microscope (SPM) system as shown in Fig. 23.22. Using an inductive sensor, the displacement $x_p(t)$ of the piezo-tube along the x-axis was measured and fed back to the DSA to compute the frequency-response (M and ϕ versus frequency ω plots). Figure 23.23 shows the Bode plots obtained by the DSA between the applied voltage $V_x(t)$ and the output of the inductive sensor $y(t)$. An estimate of the system model from the frequency-response

[12]Stanford Research Systems, model SRS785.

data was then found with the MATLAB software.[13] The transfer-function between the applied input voltage $V_x(t)$ and the output of the inductive sensor $y(t)$ was found to be

$$G_1(s) = \frac{Y(s)}{V_x(s)}$$

$$= \frac{5.544\times10^5 s^4 - 7.528\times10^9 s^3 + 1.476\times10^{15} s^2 - 4.571\times10^{18} s + 9.415\times10^{23}}{s^6 + 1.255\times10^4 s^5 + 1.632\times10^9 s^4 + 1.855\times10^{13} s^3 + 6.5\times10^{17} s^2 + 6.25\times10^{21} s + 1.378\times10^{25}} \tag{23.134}$$

with units of V/V. Equation (23.135) was scaled by the inductive sensor gain (30 Å/V) and the transfer-function between the applied voltage $V_x(t)$ and the actual displacement of the piezo-tube $x_p(t)$ is given by

$$G_2(s) = \frac{X_p(s)}{V_x(s)}$$

$$= \frac{1.663\times10^7 s^4 - 2.258\times10^{11} s^3 + 4.427\times10^{16} s^2 - 1.371\times10^{20} s + 2.825\times10^{25}}{s^6 + 1.255\times10^4 s^5 + 1.632\times10^9 s^4 + 1.855\times10^{13} s^3 + 6.5\times10^{17} s^2 + 6.25\times10^{21} s + 1.378\times10^{25}} \tag{23.135}$$

with units of Å/V.

Time Scaling of a Transfer-Function Model

We present below an approach for rescaling time for $G_2(s)$ from seconds [s] to milliseconds [ms]. We briefly recall the time scaling property of the Laplace transform presented in [1, Chapter 3, section 1.4]. Let $F(s)$ be the Laplace transform of $f(t)$, i.e.,

$$f(t) \xrightarrow{L} F(s) \tag{23.136}$$

where L denotes the Laplace transform operator. Now, consider a new time scale defined as $\hat{t} = at$, where a is a constant. The Laplace transform of $f(\hat{t}) = f(at)$ is given by

$$f(\hat{t}) = f(at) \xrightarrow{L} \frac{1}{|a|}F\left(\frac{s}{a}\right) = \hat{F}(s) \tag{23.137}$$

Using relation (23.137), we can reduce the coefficients of $G_2(s)$ by changing the time units of both the input signal $u(t)$ and output signal $y(t)$ as follows:

$$\hat{G}(s) = \frac{\hat{Y}(s)}{\hat{U}(s)} = \frac{Y(s/a)/|a|}{U(s/a)/|a|} = \frac{Y(s/a)}{U(s/a)} = G\left(\frac{s}{a}\right) \tag{23.138}$$

Therefore, to rescale time for $G_2(s)$ from seconds [s] to millisecond [ms], we choose $\hat{t} = at = 0.001t$ and the new rescaled transfer $\hat{G}_2(s)$ becomes

$$\hat{G}_2(s) = G_2\left(\frac{s}{a}\right)\Bigg|_{a=0.001}$$

$$= G_2(1000s)$$

$$\hat{G}_2(s) = \frac{16.63s^4 - 225.8s^3 + 4.427\times10^4 s^2 - 1.371\times10^5 s + 2.825\times10^7}{s^6 + 12.55s^5 + 1.632\times10^3 s^4 + 1.855\times10^4 s^3 + 6.5\times10^5 s^2 + 6.25\times10^6 s + 1.378\times10^7} \tag{23.139}$$

[13]The MATLAB command `invfreqs` gives real numerator and denominator coefficients of experimentally determined frequency response data.

Note that the time unit of the input and output signal of $\hat{G}_2(s)$ are now in milliseconds, not seconds! The coefficients of the numerator and denominator polynomials are smaller and this form ($\hat{G}_2(s)$) is less prone to computational errors due to round off than the form $G_2(s)$, Eq. (23.135).

The State-Space Model

The state-space realization for $\hat{G}_2(s)$ expressed in controllable canonical form (Eqs. (23.132) and (23.133)) is given by the following:

$$\dot{x}(t) = \begin{bmatrix} -12.55 & -1.632\times 10^3 & -1.855\times 10^4 & -6.50\times 10^5 & -6.25\times 10^6 & -1.378\times 10^7 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} x(t) + \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} u(t) \tag{23.140}$$

$$y(t) = [0 \;\; 16.63 \;\; -225.8 \;\; 4.427\times 10^4 \;\; -1.371\times 10^5 \;\; 2.825\times 10^7]x(t) \tag{23.141}$$

The time unit for Eqs. (23.140) and (23.141) are milliseconds [ms]. If the initial state at t_0 is known, along with the applied voltage $V_x(t)$ defined for $t \geq t_0$, the future behavior of the system, i.e., the state $x(t)$ and output $y(t)$, can be determined from Eqs. (23.140) and (23.141), respectively.

Discrete-Time State-Space Modeling

Introduction

The study of discrete-time systems is important to the analysis and the design of modern mechatronics systems where digital computers or small microprocessors are predominantly used to control systems. Digital computers and microprocessors output or acquire information at discrete time instants. For example, the input applied by a digital computer to actuate the piezo-tube changes at discrete time instants. Similarly, the displacment of the piezo-tube can only be measured at specified time instants using digital computers; therefore, in comparison to a continuous-time control system where the input signals change continuously over time, the input of a discrete-time system changes once in a while. Such discrete-time systems are studied next.

Consider a continuous-time system with continuous input $u(t)$ and output $y(t)$ as described by Eqs. (23.107) and (23.108). Let a digital computer or microprocessor be used to provide the input $u[k]$ and measure the output $y[k]$ as depicted in Fig. 23.24 (such systems with continuous and discrete signals are

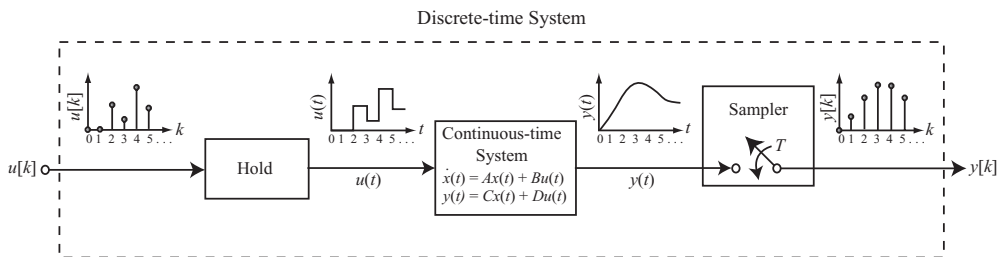



FIGURE 23.24 A block diagram of a discrete-time system showing signals in graphic from. Note that $u[k] = u(k \cdot T)$ and $y[k] = y(k \cdot T)$, for $k = 0,1,2,\ldots$, and the sampling period T is assumed to be constant.

called sampled-data systems). The input $u[k]$ and output $y[k]$ of this system are discrete with $u[k] = u(k \cdot T)$ and $y[k] = y(k \cdot T)$ for $k = 0,1,2,\ldots$, where T is the constant sampling period. The discrete input $u[k]$ is applied to the continuous system from a digital computer or microprocessor and is held constant during the time interval T (zero-order hold). A sampler acquires the output of the continuous system at each time instant T yielding the discrete output $y[k]$. The discrete system is between the input $u[k]$ and the output $y[k]$ [11, Chapter 1].[14] The equivalent discrete-time state-space representation of the continuous-time state-space model given by Eqs. (23.107) and (23.108) is given by (the details of the formulation can be found in [11, Chapter 5, section 5])

$$x[k+1] = A_D x[k] + B_D u[k] \tag{23.142}$$

$$y[k] = C_D x[k] + D_D u[k] \tag{23.143}$$

where

$$A_D = e^{AT}, \quad B_D = \left(\int_0^T e^{A\lambda} d\lambda\right) B, \quad C_D = C, \text{ and } D_D = D \tag{23.144}$$

and matrices C_D and D_D are not changed by the sampling.[15] This discrete model (Eqs. (23.142) and (23.143)) is the representation of the sampled-data system shown in Fig. 23.24.

Solutions to the Discrete-Time State-Space Equations

The solution to the discrete model (Eqs. (23.142) and (23.143)) is given by

$$x[k] = A_D^k x[0] + \sum_{j=0}^{k-1} A_D^{k-j-1} B_D u[j] \tag{23.145}$$

$$y[k] = CA_D^k x[0] + C\sum_{j=0}^{k-1} A_D^{k-j-1} B_D u[j] + Du[k] \tag{23.146}$$

for each sampling step k. Details of the formulation can be found in [11, Chapter 5, section 3]. The state response $x[k]$ to an applied input $u[k]$ is characterized by the system matrices (A_D, B_D, C_D, D_D). In particular, the output $y[k]$ will be bounded for any bounded input $u[k]$ if the system is stable. A system in the form given by Eq. (23.142) is stable if the magnitude of all the eigenvalues of A_D are less than unity, i.e., lie within the unit circle center at the origin of the z-plane [11, Chapter 5, section 6].

The z-Transform and Relationship with the State-Space

The input-to-output relationship in the frequency-domain for a discrete-time system is represented by a discrete transfer-function called the z-transform, written in terms of the variable z [12, Chapter 4]. Analogous to the continuous-time case, the model of a dynamic system in discrete transfer-function form can be useful in the design and control of systems [12, Chapter 7]. If the system model is available in discrete transfer-function form, then a state-space realization can be found as follows. Given a discrete system described by the following z-transform $G(z)$:

$$G(z) = \frac{d_0 + d_1 z^{-1} + \cdots + d_n z^{-n}}{1 + c_1 z^{-1} + \cdots + c_n z^{-n}} \tag{23.147}$$

[14] We do not discuss quantizing and quantization error. See [11, Chapter 1, section 3] for details.

[15] Given a continuous-time state-space model (A, B, C, D), the MATLAB command `c2d`, gives the discrete time equivalent for a specified sampling period T.

the controllable canonical discrete state-space realization for $G(z)$ is

$$x[k+1] = \begin{bmatrix} -c_1 & -c_2 & \cdots & -c_{n-1} & -c_n \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix} x[k] + \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} u[t] \tag{23.148}$$

$$y[k] = [(d_1 - c_1 d_0) \;\; (d_2 - c_2 d_0) \;\; \ldots \;\; (d_n - c_n d_0)]x[k] + [d_0]u[k] \tag{23.149}$$

The number of states n is equivalent to the highest power of the denominator of $G(z)$. For information about other equivalent canonical state-space forms, refer to [11, Chapter 5, section 2].

Example

Consider the continuous-time state-space model of the piezo-tube system described by Eqs. (23.140) and (23.141). A digital computer with the sampling rate of 10 kHz ($T = 1.0 \times 10^{-4}$) is used to provide the control input $u[k]$ and measure its displacement along the x-axis (output $y[k]$). The discrete-time state-space model with (A_D, B_D, C_D, and D_D) given by Eq. (23.144) is

$$x[k+1] = \begin{bmatrix} 0.999 & -0.163 & -1.85 & -65.0 & -624.5 & -1377.1 \\ 9.99\times10^{-5} & 0.999 & -9.26\times10^{-5} & -3.25\times10^{-3} & -3.12\times10^{-2} & -6.69\times10^{-2} \\ 5.00\times10^{-9} & 1.00\times10^{-4} & 1 & -1.08\times10^{-7} & -1.04\times10^{-6} & -2.30\times10^{-6} \\ 1.67\times10^{-13} & 5.00\times10^{-9} & 1.00\times10^{-4} & 1 & -2.60\times10^{-11} & -5.74\times10^{-11} \\ 4.17\times10^{-18} & 1.67\times10^{-13} & 5.00\times10^{-9} & 1.00\times10^{-4} & 1 & -1.15\times10^{-15} \\ 8.33\times10^{-23} & 4.17\times10^{-18} & 1.67\times10^{-13} & 5.00\times10^{-9} & 1.00\times10^{-4} & 1 \end{bmatrix} x[k]$$

$$+ \begin{bmatrix} 9.99\times10^{-5} \\ 4.99\times10^{-9} \\ 1.67\times10^{-13} \\ 4.17\times10^{-18} \\ 8.33\times10^{-23} \\ 1.39\times10^{-27} \end{bmatrix} u[k] \tag{23.150}$$

$$y[k] = \lfloor 0 \;\; 16.63 \;\; -225.8 \;\; 4.427\times10^{4} \;\; -1.371\times10^{5} \;\; 2.825\times10^{7} \rfloor x[k] \tag{23.151}$$

The realization given by Eqs. (23.150) and (23.151) was found using the MATLAB command '`c2d`'.

Summary

We presented tools for modeling continuous- and discrete-time systems using the state-space approach in this section. The state-space approach to modeling is a powerful technique for the analysis and design of mechatronic and dynamic systems, and can take advantage of tools available in modern digital computers and microprocessors. The discussion of the system states and the state-space was motivated by an example piezo-tube actuator system. We considered the modeling of linear systems and a technique for linearizing nonlinear systems was briefly introduced. The frequency-response of a system and an approach to modeling using experimental frequency-response data was presented. Relationships between models

expressed in the frequency- and time-domain for both continuous- and discrete-time systems was discussed. For additional details about the concepts mentioned in this section and those not covered, it is recommended that the reader consider the attached references for further reading.

References

1. Franklin, G. F., et al., *Feedback Control of Dynamic Systems,* 3rd ed., Addison-Wesley, New York, 1994.
2. Hanselman, D., and Littlefield, B., *The Student Edition of Matlab, Version 5, User's Guide,* Prentice-Hall, Upper Saddle River, NJ, 1997.
3. Croft, D., et al., Creep, hysteresis, and vibration compensation for piezoactuators: atomic force microscopy application, *ASME J. Dyn. Syst., Meas., Control,* 123, 35, 2001.
4. Dorny, C. N., *Understanding Dynamic Systems—Approaches to Modeling, Analysis, and Design,* Prentice-Hall, Englewood Cliffs, NJ, 1993.
5. Locatelli, M., et al., Easy method to characterize a piezoelectric ceramic tube as a displacer, *Rev. Sci. Instrum.,* 59, 4, 1988.
6. Dorf, R. C., and Bishop, R. H., *Modern Control Systems,* 9th ed., Prentice-Hall, Upper Saddle River, NJ, 2001.
7. Chen, T. C., *Linear System Theory and Design,* Oxford University Press, New York, 1999.
8. Friedland, B., *Control System Design: An Introduction to State-Space Methods,* McGraw-Hill, New York, 1986.
9. Gillis, J. T., State space, in T*he Control Handbook,* Levine, W. S., CRC Press, Salem, MA, 1996, Chap. 5.
10. Khalil, H. K., *Nonlinear Systems,* 2nd ed., Prentice-Hall, Upper Saddle River, NJ, 1996.
11. Ogata, K., *Discrete-Time Control Systems,* 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 1995.
12. Franklin, G. F., et al., *Digital Control of Dynamic Systems,* 3rd ed., Addison-Wesley, Menlo Park, 1998.

23.4 Transfer Functions and Laplace Transforms

C. Nelson Dorny

We perceive a system primarily through its behavior. Therefore, our mental image of a system usually includes representative response **signals**. The *step response*, the behavior when we suddenly turn on the system, is such a system-characterizing signal. We should view the step response as a description of the system. The *impulse response* is another description of the system. For a system represented by linear differential equations, the unit-step response is the integral of the unit-impulse response.

Let us represent time differentiation (d/dt) by the *time-derivative operator, p*. Then we can denote the time derivative of a signal y by py, its second derivative by p^2y, its integral with respect to time by $(1/p)y$, and so on. This *operator notation* simplifies the expressions for differential equations. We shall use the expression *system equations* to mean a set of differential equations that determines fully the behaviors of the dependent variables that appear in those equations. We can reduce a *linear* set of system equations to a single **input–output system equation** by eliminating all but one dependent variable from the set. The *transfer function* associated with that dependent variable is a mathematical expression that contains all the essential information embodied in the system differential equation.

The Laplace transformation converts signals (functions of time) to functions of a *complex-frequency variable,* $s = \sigma + j\omega$. There is a one-to-one correspondence between a signal and its Laplace transform. We can retrieve the time function by inverse transformation. Laplace transformation produces images that have some properties that are more convenient than those of the original signals. In particular, time differentiating a signal corresponds to multiplying its Laplace transform by the complex-frequency variable s. Hence, the transformation converts linear constant-coefficient differential equations to linear algebraic equations. Such simplifications of time-domain operations make Laplace transformation useful.

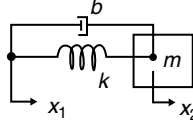



FIGURE 23.25 The lumped model of a mechanical system.

The Laplace transformation also converts the impulse response of a system variable to the transfer function for that variable. As a consequence, we can view the differential equation that represents a linear system as an expression of the response of that system to an impulsive input.

Transfer Functions

The node displacements x_1 and x_2 and the compressive forces f_1 and f_2 within the branches of the lumped model of Fig. 23.25 are related to each other by the spring equation, the damper equation, and the balance of forces at node 2. The spring equation is $f_2 = k(x_1 - x_2)$. The equation for the damper is $f_1 = b(px_1 - px_2)$. The balance of forces requires that $f_1 + f_2 = mp^2x_2$. These equations describe fully the behavior of the system if the spring and mass are unenergized. (If the mass were moving and/or the spring were compressed, we would have to express separately their initial energy states to describe fully the future relations among the variables.)

Eliminate f_1 and f_2 from the equations to obtain the operational equation:

$$(mp^2 + bp + k)x_2 = (bp + k)x_1 \tag{23.152}$$

This differential equation describes fully the **zero-state** relation between x_1 and x_2. Rearrange Eq. (23.152) to form the ratio

$$\frac{x_2}{x_1} = \frac{bp + k}{mp^2 + bp + k} \tag{23.153}$$

We call Eq. (23.152) the *transfer function* from x_1 to x_2. The transfer function focuses attention on the mathematical operations that characterize the behavioral relationships rather than on the particular natures of the variables. (Note that the transfer function from v_1 to v_2, where $v_1 = px_1$ and $v_2 = px_2$, is the same as the transfer function given by Eq. (23.153).)

In general, suppose that y_1 and y_2 are two variables related (in operator notation) by the linear differential equation

$$y_2 = G(p)y_1 \tag{23.154}$$

We formally define the *transfer function* from y_1 to y_2 by

$$G(p) = \left.\frac{y_2}{y_1}\right|_{\text{ZS}} \tag{23.155}$$

where the notation ZS means **zero state**. If y_1 is an independent variable, then $G(p)$ is the *input–output transfer function* for the variable y_2 and accounts fully for its behavior owing to the **input signal** y_1. We can determine from that transfer function the behavior of the system for any source waveform and any initial state.

The Laplace Transformation

The *one-sided Laplace transformation*, $\mathscr{L}$, is an integral operator that converts a signal $f(t)$ to a complex-valued function $F(s)$ in the following fashion:

$$\mathscr{L}[f(t)] \equiv F(s) \triangleq \int_{0^-}^{\infty} f(t)e^{-st}\,dt \tag{23.156}$$

We refer to the transformed function $F(s)$ as the *Laplace transform* of the signal $f(t)$. Picture the lower limit 0^- of the integral as a *specific* instant prior to but infinitesimally close to $t = 0$. It is customary to use a lowercase symbol (f) to represent a signal waveform and an uppercase symbol (F) to represent its Laplace transform. (Although we speak here of time signals, there is nothing in Eq. (23.156) that requires $f(t)$ to be a function of time. The transformation can be applied to functions of any quantity t.)

We shall use the Laplace transformation to transform the signals of **time-invariant** linear systems. The behavior of such a system for $t \geq 0$ depends only on the input signal for $t \geq 0$ and on the prior **state** of the output variable (at $t = 0^-$). Hence, it does not matter that the Laplace transformation ignores $f(t)$ for $t < 0^-$.

The process of finding the time function $f(t)$ that corresponds to a particular Laplace transform $F(s)$ is called *inverse Laplace transformation*, and is denoted by $\mathscr{L}^{-1}$. We also call $f(t)$ the *inverse Laplace transform* of $F(s)$. Since the one-sided Laplace transformation ignores $t < 0^-$, $F(s)$ contains no information about $f(t)$ for $t < 0^-$. Therefore, inverse Laplace transformation cannot reconstruct $f(t)$ for $t < 0^-$. We shall treat all signals as if they are defined only for $t \geq 0^-$. Then there is a one-to-one relation between $f(t)$ and $F(s)$.

To illustrate the Laplace transformation, we find the Laplace transform of the decaying exponential, $f(t) = e^{-\alpha t}$, $t \geq 0^-$. The transform is

$$\begin{aligned} F(s) &= \int_{0^-}^{\infty} e^{-\alpha t}e^{-st}\,dt = \left.\frac{e^{-(s+\alpha)t}}{-(s+\alpha)}\right|_{0^-}^{\infty} \\ &= \left.\frac{e^{-(\sigma+\alpha)t}e^{-j\omega t}}{-(s+\alpha)}\right|_{0^-}^{\infty} = \frac{1}{s+\alpha} \quad \text{for } \mathrm{Re}[s] > -\alpha \end{aligned} \tag{23.157}$$

We must require $\sigma > -\alpha$, where σ is the real part of s, in order that the real-exponent factor converge to zero at the upper limit. (The magnitude of the complex-exponent factor remains 1 for all t.) Therefore, the Laplace transform of the decaying exponential is defined only for $\mathrm{Re}[s] > -\alpha$. This restriction on the domain of F in the complex s plane is comparable to the restriction $t \geq 0^-$ on the domain of f.

The significant features of the complex-frequency function $1/(s + \alpha)$ are the existence of a single pole and the location of that pole, $s = -\alpha$ [rad/s]. (The pole defines the left boundary of that region of the complex s plane over which the transform $1/(s + \alpha)$ is defined.) The significant features of the corresponding time function are the fact of decay and the rate of decay, with the exponent $-\alpha$ [rad/s]. There are clear parallels between the features of $f(t)$ and $F(s)$. We should think of the whole complex-valued function F as representing the whole time waveform f.

As a second transformation example, let $f(t) = \delta(t)$, the unit impulse, essentially a unit-area pulse of very short duration. It acts at $t = 0$, barely within the lower limit of the Laplace integral. It has value zero at $t = 0^-$. (Because we use 0^- as the lower limit of the defining integral, it does not matter whether the impulse straddles $t = 0$ or begins to rise at $t = 0$.) The impulse is nonzero only for $t \approx 0$, where $e^{-st} \approx 1$. Therefore, the Laplace transform is

$$\Delta(s) = \int_{0^-}^{\infty} \delta(t)e^{-st}\,dt \approx \int_{0^-}^{\infty} \delta(t)(1)\,dt = 1 \tag{23.158}$$

Table 23.11 Laplace Transform Pairs

$f(t) = \mathscr{L}^{-1}[\mathbf{F}(s)], t \geq 0^-$	$\mathbf{F}(s) = \mathscr{L}[f(t)]$
1. Unit impulse $\delta(t)$	1
2. Unit step $u_s(t)$	$\frac{1}{s}$
3. t^n, $n = 1, 2, \ldots$	$\frac{n!}{s^{n+1}}$
4. $e^{-\alpha t}$	$\frac{1}{s+\alpha}$
5. $t^n e^{-\alpha t}$, $n = 1, 2, \ldots$	$\frac{n!}{(s+\alpha)^{n+1}}$
6. $\sin(\omega_0 t)$	$\frac{\omega_0}{s^2 + \omega_0^2}$
7. $\cos(\omega_0 t)$	$\frac{s}{s^2 + \omega_0^2}$
8. $e^{-\alpha t}\sin(\omega_d t)$	$\frac{\omega_d}{(s+\alpha)^2 + \omega_d^2}$
9. $e^{-\alpha t}\cos(\omega_d t)$	$\frac{s+\alpha}{(s+\alpha)^2 + \omega_d^2}$

Source: Dorny, C. N. 1993. *Understanding Dynamic Systems,* p. 412. Prentice-Hall, Englewood Cliffs, NJ. With permission.

It is not necessary to derive the Laplace transform for each signal that we use in the study of systems. Table 23.11 gives the transforms for some signal waveforms that are common in dynamic systems.

Transform Properties

A number of useful properties of the Laplace transformation $\mathscr{L}$ are summarized in Table 23.12. According to the derivative property, the multiplier s acts precisely like the time-derivative operator, but in the domain of Laplace-transformed signals. When we Laplace transform the equation for an energy-storage element such as a mass or a spring, the derivative property automatically incorporates the prior energy

Table 23.12 Properties of the Laplace Transformation, $\mathscr{L}$

1. Magnification	$\mathscr{L}[af(t)] = a\mathbf{F}(s)$
2. Addition	$\mathscr{L}[f_1(t) + f_2(t)] = \mathbf{F}_1(s) + \mathbf{F}_2(s)$
3. Derivative	$\mathscr{L}[\dot{f}(t)] = s\mathbf{F}(s) - f(0^-)$
4. Derivatives	$\mathscr{L}[\ddot{f}(t)] = s^2\mathbf{F}(s) - sf(0^-) - \dot{f}(0^-)$
5. Integral	$\mathscr{L}\left[\int_{0^-}^{t} f(t)\,dt\right] = \frac{\mathbf{F}(s)}{s}$
6. Convolution	$\mathscr{L}\left[\int_{0^-}^{t} f_1(\lambda)f_2(t-\lambda)\,d\lambda\right] = \mathbf{F}_1(s)\mathbf{F}_2(s)$
7. Initial value	$f(0^+) = \lim_{t\to 0^+} f(t) = \lim_{s\to\infty} s\mathbf{F}(s)$
8. Final Value	$f(\infty) = \lim_{t\to\infty} f(t) = \lim_{s\to 0} s\mathbf{F}(s)$ if finite
9. Definite integral	$\int_0^\infty f(t)\,dt = \lim_{s\to 0} s\mathbf{F}(s)$ if finite
10. Exponential decay	$\mathscr{L}[e^{-\alpha t} f(t)] = \mathrm{F}(s+\alpha)$
11. Delay	$\mathscr{L}[f(t-t_0)u_s(t-t_0)] = e^{-t_0 s}\mathbf{F}(s)$ for $t_0 \geq 0$
12.Time multiplication	$\mathscr{L}[tf(t)] = -\frac{d\mathbf{F}(s)}{ds}$
13. Time division	$\mathscr{L}\left[\frac{f(t)}{t}\right] = \int_s^\infty \mathbf{F}(s)\,ds$
14. Time scaling	$\mathscr{L}[f(at)] = \frac{\mathbf{F}(s/\mathrm{a})}{a}$

Source: Dorny, C. N. 1993. *Understanding Dynamic Systems*, p. 413. Prentice-Hall, Englewood Cliffs, NJ. With permission.

state of the element—essentially the value of the variable at $t = 0^-$. When we Laplace transform the input–output system equation for a particular system variable, the derivative property automatically incorporates the whole prior system state. As a consequence, we can find the solution to the system equation without having to determine the initial conditions (at $t = 0^+$)—a considerable simplification of the solution process.

Since $F(s)$ contains all information about $f(t)$ for $t \geq 0^-$, it is possible to find some features of the signal $f(t)$ from the transform $F(s)$ without performing an inverse Laplace transformation. Properties 7–9 of Table 23.12 provide three of these features, namely the initial value ($t \to 0^+$), the final value ($t \to \infty$), and the area under the waveform. The remaining properties in the table show the effect on the transform of various changes in the signal waveform.

The usual approach to finding inverse transforms is to use a table of transform pairs. That table might be stored in a software package such as CC, MATLAB, MAPLE, and so on. Table 23.11 demonstrates that transforms of typical system signals are ratios of polynomials in s. A ratio of polynomials can be decomposed into a sum of *simple* polynomial fractions—a process referred to as *partial fraction expansion*. Hence, the inversion process can be accomplished by a computer program that incorporates a brief table of transforms.

Transformation and Solution of a System Equation

Suppose that an independent external source applies a specific velocity pattern $v_1(t)$ to node 1 of Fig. 23.25. To obtain the input–output system equation that relates the velocity v_2 of node 2 to the input signal v_1, multiply Eq. (23.153) by p and substitute v_1 for px_1 and v_2 for px_2. The result is

$$(mp^2 + bp + k)v_2 = (bp + k)v_1 \tag{23.159}$$

The two sides of Eq. (23.159) are identical functions of time. Therefore, the Laplace transforms of the two sides of Eq. (23.159) are equal. Since the Laplace transformation is linear (properties 1 and 2 of Table 23.12), and since the coefficients of the differential equation are constants, the Laplace transform can be applied separately to the individual terms of each side. The result is

$$m[s^2V_2(s) - sv_2(0^-) - \dot{v}_2(0^-)] + b[sV_2(s) - v_2(0^-)] + kV_2(s) = b[sV_1(s) - v_1(0^-)] + kV_1(s) \tag{23.160}$$

where the derivative properties of the Laplace transformation (properties 3 and 4 of Table 23.12) introduce the prior values $v_1(0^-)$, $v_2(0^-)$, and $\dot{v}_2(0^-)$ into the equation. According to Eq. (23.160), to fully determine the transform $V_2(s)$ of the behavior $v_2(t)$, we must specify these prior values and also $V_1(s)$. It can be shown that specifying the three prior values is equivalent to specifying the energy states of the spring and mass.

Let us assume that the independent source applies the constant velocity $v_1(t) = v_c$ beginning at $t = 0$. The corresponding transform, by item 2 of Table 23.11 and property 1 of Table 23.12, is $V_1(s) = v_c/s$. Substitute the transform $V_1(s)$ into Eq. (23.160) and solve for

$$V_2(s) = \frac{(bs+k)v_c + ms\dot{v}_2(0^-) + bs[v_2(0^-) - v_1(0^-)] + ms^2v_2(0^-)}{s(ms^2+bs+k)} \tag{23.161}$$

We could find the output signal waveform $v_2(t)$ as a function of the model parameters m, k, b, the source-signal parameter v_c, and the prior state information $v_1(0^-)$, $v_2(0^-)$, and $\dot{v}_2(0^-)$, but the expression for the solution would be messy. Instead, we complete the solution process for specific numbers: $m = 2$ kg, $b = 4$ N · s/m, $k = 10$ N/m, $\dot{v}_2(0^-) = 0$ m/s^2, $v_1(0^-) = 0$ m/s, $v_2(0^-) = -1$ m/s, and $v_c = 1$ m/s. The partial-fraction expansion of the transform and the inverse transform, both obtained by a commercial computer program, are

$$V_2(s) = \frac{1}{s} - \frac{2s+2}{(s+1)^2+2^2} \tag{23.162}$$

$$v_2(t) = 1 - 2e^{-t}\cos(2t), \quad \text{for } t \geq 0 \tag{23.163}$$

We can take Laplace transforms of the system equations at any stage in their development. We can even write the equations directly in terms of transformed variables if we wish. The process of eliminating variables can be carried out as well in one notation as in another. For example, the operator $G(p)$ in Eq. (23.154) represents a ratio of polynomials in the time-derivative operator p. Therefore, Laplace transforming the differential equation, Eq. (23.154), introduces the prior values of various derivatives of y_1 and y_2. If the prior values of all these derivatives are zero, then the Laplace-transformed equation is

$$Y_2(s) = G(s)Y_1(s) \tag{23.164}$$

where the operator p in Eq. (23.154) is replaced by the complex-frequency variable s in Eq. (23.164). It is appropriate, therefore, to define the transfer function directly in terms of Laplace-transformed signals:

$$G(s) = \left.\frac{Y_2(s)}{Y_1(s)}\right|_{\text{PV}=0} \tag{23.165}$$

where $Y_1(s)$ and $Y_2(s)$ are the Laplace transforms of the signals $y_1(t)$ and $y_2(t)$, and the notation PV = 0 means that the prior values (at $t = 0^-$) of $y_1(t)$ and $y_2(t)$ and the various derivatives mentioned above in connection with Eq. (23.164) are set to zero. The *frequency domain* definition, Eq. (23.165), is equivalent to the *time domain* definition, Eq. (23.155).

Suppose that the input signal $y_1(t)$ is the unit impulse $\delta(t)$. Then the response signal $y_2(t)$ is the unit-impulse response of the system. Since the Laplace transform of the unit impulse is $Y_1(s) = \Delta(s) = 1$ by entry 1 of Table 23.11, Eq. (23.164) shows that the Laplace transform $Y_2(s)$ of the unit-impulse response is identical to the zero-state transfer function (expressed in the transform domain).

The transfer function for a linear system has two interpretations. Both interpretations characterize the system. In the frequency domain, the transfer function $G(s)$ is the multiplier that produces the response—by multiplying the source-signal transform, as in Eq. (23.164). In the time domain, we use a representative response signal—the impulse response—to characterize the system. The transfer function $G(s)$ is the Laplace transform of that characteristic response.

Defining Terms

Input: An independent variable.

Input–output system equation: A differential equation that describes the behavior of a single dependent variable as a function of time. The dependent variable is viewed as the system output. The independent variable(s) are the inputs.

Output: A dependent variable.

Signal: An observable variable; a quantity that reveals the behavior of a system.

State: The state of an nth-order linear system corresponds to the values of a dependent variable and its first $n - 1$ time derivatives.

Time invariant: A system that can be represented by differential equations with constant coefficients.

Zero state: A condition in which no energy is stored or in which all variables have the value zero.

References

Franklin, G. F., Powell, J. D., and Emami-Naeini, A. 1994. *Feedback Control of Dynamic Systems,* 3rd ed., Addison Wesley, Reading, MA.

Kuo, B. C. 1991. *Automatic Control Systems,* 6th ed., Prentice-Hall, Englewood Cliffs, NJ.

Nise, N. S. 1992. *Control Systems Engineering,* Benjamin Cumming, Redwood City, CA.

Further Information

A thorough mathematical treatment of Laplace transforms is presented in *Advanced Engineering Mathematics,* by C. Ray Wylie and Louis C. Barrett. *Understanding Dynamic Systems,* by C. Nelson Dorny, applies transfer functions and related concepts in a variety of contexts. The following journals publish papers that use transfer functions and Laplace transforms:

IEEE Transactions on Automatic Control. Published monthly by the Institute of Electrical and Electronics Engineers.

IEEE Transactions on Systems, Man, and Cybernetics. Published bimonthly.

Journal of Dynamic Systems, Measurement, and Control. Published quarterly by the American Society of Mechanical Engineers.

24

State Space Analysis and System Properties

Mario E. Salgado
Universidad Técnica Federico Santa María

Juan I. Yuz
Universidad Técnica Federico Santa María

24.1 Models: Fundamental Concepts

An essential connection between an engineer/scientist and a system relies on his/her ability to describe the system in a way which is useful to understand and to quantify its behavior.

Any description supporting that connection is a **model**. In system theory, models play a fundamental role, since they are needed to analyze, to synthesize, and to design systems of all imaginable sorts.

There is not a unique model for a given system. Firstly, the need for a model may obey different purposes. For instance, when dealing with an electric motor, we might be interested in the electro-mechanical energy conversion process, alternatively, we might be interested in modelling the motor either as a thermal system, or as a mechanical system to study vibrations, the strength of the materials, and so on.

A second source of that nonuniqueness is the fact that models are always inaccurate, since real systems are usually infinitely complex. One of the key decisions for an engineer when facing the task of **modelling** a system is to decide which are the essential features that the model should capture, and that decision is also closely related to the purpose of the model.

The theory supporting modelling is by itself a vast field, where first principles, signal theory, mathematics and numerical tools combine in different ways to generate rich methodologies. A model is rarely built in one go, the model building process is usually iterative, and it progresses according to the quality of the results obtained when using the model in a particular application. Iterations may also include changes in modelling methodology.

In this chapter we will deal with a special class of models to describe *dynamic systems*. Dynamic systems are those where the system variables are interdependent not only algebraically, but also in a way where we observe the intervention of accumulated effects and rate of change. Models for dynamic systems can be built in the continuous time domain, in the discrete time domain, or in a continuous-discrete time framework (for hybrid systems, involving sampled systems). We will cover the three situations.

In this quest we will put the emphasis on concepts, fundamental properties, physical interpretations, and examples. We will include neither proofs nor intricate theoretical developments. Sometimes we will sacrifice rigor for the sake of an easier understanding. To cover in depth the theory supporting our presentation we refer the interested reader to the specialized literature such as [6,8,10–14].

24.2 State Variables: Basic Concepts

Introduction

One of the most frequently used class of models is that defined by a set of equations on a set of system inner variables. These inner variables are known as **state variables**. The values they have at a specific time instant form a set known as the **system state**, although we will often use the expressions *state variables* and *system state* as synonyms.

The above definition is too vague since it would fit to any set of system variables. What is distinctive in the set of state variables is clarified in the following definition.

A set of state variables for the given system is a set of system inner variables such that any system variable can be computed as a function of the present state and the present and future system inputs.

In this definition we have preferred to stress the physical meaning of state variables. However, a more abstract definition is also possible. The definition also implies that if we know the state at time t we can then compute the energy stored in the system at that instant. The energy stored in a system depends on some system variables (speed, voltage, current, position, temperature, pressure, etc.) and all of them, by definition, can be computed from the system state.

The above definition suggests that one can think of the state in a more general way: the state variables can be chosen as a **function** (e.g., a linear combination) of inner system variables. This generalization builds some distance between the state and its physical interpretation. However, it has the advantage of making the framework more general. It also makes more evident an interesting feature: the **choice of state variables is not unique**.

Another important observation is that the time evolution of the state, the state trajectory itself, can be computed from the present value of the state and the present and future inputs. Thus, the models involved are first order differential (continuous time) or one-step recursive (discrete time) equations.

Basic State Space Models

If we denote by $\mathbf{x}$ the vector corresponding to a particular choice of state variables, the general form of a state variable model is as follows:

For continuous-time systems

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x}(t), \mathbf{u}(t), t) \tag{24.1}$$

$$\mathbf{y}(t) = \mathbf{G}(\mathbf{x}(t), \mathbf{u}(t), t) \; (24.2)$$

where $\mathbf{u}(t)$ is the system **input vector** and $\mathbf{y}(t)$ is an **output vector**.

For discrete-time systems

$$\mathbf{x}[t+1] = \mathbf{F_d}(\mathbf{x}[t], \mathbf{u}[t], t) \tag{24.3}$$

$$y[t] = \mathbf{G_d}(\mathbf{x}[t], \mathbf{u}[t], t) \tag{24.4}$$

Similarly to the continuous-time case, $\mathbf{u}[t]$ is the system **input vector** and $\mathbf{y}[t]$ is an **output vector**.

Note that throughout this chapter we will use the symbol t to denote continuous and discrete time, but the difference will be made on using [and] to enclose the argument in the discrete-time case, when t Œ $\mathbb{Z}$.

To obtain a first glimpse at the concepts underlying the state space approach, we consider the following example.

Example 24.1

In Fig. 24.1, an external force $f(t)$ is applied to a mass-spring system. The position $d(t)$ is measured with respect to the mass position when the spring is relaxed and no external force is applied. The mass movement is damped by a viscous friction force proportional to the mass velocity, $v(t)$.

From first principles we know that to be able to compute the mass position and the mass velocity we must know the initial mass velocity, and the initial spring stretching. Thus, the state vector must have two components, i.e., $\mathbf{x}(t) = [x_1(t) \quad x_2(t)]^T$, and a natural state choice is

$$x_1(t) = d(t) \; (24.5)$$

$$x_2(t) = v(t) = \dot{x}_1(t) \tag{25.6}$$

With this choice, one can apply Newton laws to obtain

$$f(t) = M\frac{dv(t)}{dt} + Kd(t) + Dv(t) = M\dot{x}_2(t) + Kx_1(t) + Dx_2(t) \tag{24.7}$$

where D is the viscous friction proportional constant. We are now in position to write the state equations as

$$\dot{x}_1(t) = x_2(t) \; (24.8)$$

$$\dot{x}_2(t) = -\frac{K}{M}x_1(t) - \frac{D}{M}x_2(t) + \frac{1}{M}f(t) \; (24.9)$$

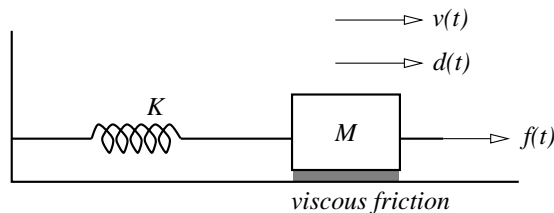



FIGURE 24.1 Mechanical system.

We also observe that the energy, $w(t)$, stored in the system is given by

$$w(t) = \frac{1}{2}K(d(t))^2 + \frac{1}{2}M(v(t))^2 = \mathbf{x}(t)^T\mathbf{\Lambda}\mathbf{x}(t) \tag{24.10}$$

where $\mathbf{\Lambda}$ is a diagonal matrix: $\mathbf{\Lambda} = \mathrm{diag}\left\{\frac{K}{2}, \frac{M}{2}\right\}$.

Finally, the nonuniqueness of the state vector can be appreciated if, instead of the choices made in (24.8), we choose a new state $\bar{\mathbf{x}}(t)$ related to $\mathbf{x}(t)$ by a nonsingular matrix $\mathbf{T} \in \mathbb{R}^{2\times 2}$, i.e.,

$$\bar{\mathbf{x}}(t) = \mathbf{T}\mathbf{x}(t) \tag{24.11}$$

More on this will be said in subsection "State Similarity Transformation."

Signals and State Space Description

The state space framework can also be used to describe a wide variety of signals using a model of the form

$$\frac{d\,\mathbf{x}(t)}{dt} = \mathbf{A}\mathbf{x}(t), \qquad y(t) = \mathbf{C}\mathbf{x}(t) \quad \text{for continuous-time signals} \tag{24.12}$$

$$\mathbf{x}[t+1] = \mathbf{A}_q\mathbf{x}[t], \qquad y[t] = \mathbf{C}_q\mathbf{x}[t] \quad \text{for discrete-time signals} \tag{24.13}$$

To illustrate the idea we consider a continuous-time signal given by

$$f(t) = 2 + 4\cos(5t) - \sin(5t) \tag{24.14}$$

This signal can be interpreted as the solution for the homogeneous differential equation

$$\frac{d^3 f(t)}{dt^3} + 25\frac{df(t)}{dt} = 0, \quad \text{subject to } f(0) = 6,\ \dot{f}(0) = -5 \ \text{ and } \ \ddot{f}(0) = -100 \tag{24.15}$$

If we now choose, as state variables, $x_1(t) = f(t)$, $x_2(t) = \dot{f}(t)$, and $x_3(t) = \ddot{f}(t)$, then the state space model for this signal is

$$\frac{d\mathbf{x}(t)}{dt} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -25 & 0 \end{bmatrix}\mathbf{x}(t), \qquad y(t) = [1 \quad 0 \quad 0]\,\mathbf{x}(t) \tag{24.16}$$

In this usage of state space models, the state variables have no particular physical meaning. However, this description is particularly useful in signal reconstruction theory and when dealing with disturbances in control system synthesis.

24.3 State Space Description for Continuous-Time Systems

In this section the state space description for continuous-time systems is presented. The analysis is focused on the class of **linear and time invariant** systems; to do that, we first show how to build a linear model from the nonlinear equations (24.1) and (24.2).

An additional restriction is that, at this stage, the systems under study have no pure time delays. This feature generates an infinite dimensional state vector. However, we will see in section 24.4 that this class of systems can be successfully dealt with using sampled data models.

Linearization

Since we will concentrate on time invariant systems, (24.1) and (24.2) can be rewritten as

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x}(t), \mathbf{u}(t)) \tag{24.17}$$

$$\mathbf{y}(t) = \mathbf{G}(\mathbf{x}(t), \mathbf{u}(t)) \tag{24.18}$$

We assume that the model (24.17) and (24.18) has at least one *equilibrium point* given by $\{\mathbf{x}_Q, \mathbf{u}_Q, \mathbf{y}_Q\}$. This is a triad conformed by three constant vectors satisfying

$$\mathbf{0} = \mathbf{F}(\mathbf{x}_Q, \mathbf{u}_Q) \tag{24.19}$$

$$\mathbf{y}_Q = \mathbf{G}(\mathbf{x}_Q, \mathbf{u}_Q) \tag{24.20}$$

Note that the equilibrium point is defined by the state derivatives equal to zero.

If we now consider a neighborhood around the equilibrium point, then we can approximate the model (24.17) and (24.18) by a truncated Taylor's series having the form

$$\dot{\mathbf{x}}(t) \approx \mathbf{F}(\mathbf{x}_Q, \mathbf{u}_Q) + \left.\frac{\partial \mathbf{F}}{\partial \mathbf{x}}\right|_{\substack{\mathbf{x}=\mathbf{x}_Q\\ \mathbf{u}=\mathbf{u}_Q}}(\mathbf{x}(t) - \mathbf{x}_Q) + \left.\frac{\partial \mathbf{F}}{\partial \mathbf{u}}\right|_{\substack{\mathbf{x}=\mathbf{x}_Q\\ \mathbf{u}=\mathbf{u}_Q}}(\mathbf{u}(t) - \mathbf{u}_Q) \tag{24.21}$$

$$\mathbf{y}(t) \approx \mathbf{G}(\mathbf{x}_Q, \mathbf{u}_Q) + \left.\frac{\partial \mathbf{G}}{\partial \mathbf{x}}\right|_{\substack{\mathbf{x}=\mathbf{x}_Q\\ \mathbf{u}=\mathbf{u}_Q}}(\mathbf{x}(t) - \mathbf{x}_Q) + \left.\frac{\partial \mathbf{G}}{\partial \mathbf{u}}\right|_{\substack{\mathbf{x}=\mathbf{x}_Q\\ \mathbf{u}=\mathbf{u}_Q}}(\mathbf{u}(t) - \mathbf{u}_Q) \tag{24.22}$$

Equations (24.21) and (24.22) can then be written as

$$\frac{d\Delta\mathbf{x}(t)}{dt} = \mathbf{A}\Delta\mathbf{x}(t) + \mathbf{B}\Delta\mathbf{u}(t) \tag{24.23}$$

$$\Delta\mathbf{y}(t) = \mathbf{C}\Delta\mathbf{x}(t) + \mathbf{D}\Delta\mathbf{u}(t) \tag{24.24}$$

where

$$\Delta\mathbf{x}(t) = \mathbf{x}(t) - \mathbf{x}_Q, \qquad \Delta\mathbf{u}(t) = \mathbf{u}(t) - \mathbf{u}_Q, \qquad \Delta\mathbf{y}(t) = \mathbf{y}(t) - \mathbf{y}_Q \tag{24.25}$$

and

$$\mathbf{A} = \left.\frac{\partial \mathbf{F}}{\partial \mathbf{x}}\right|_{\substack{\mathbf{x}=\mathbf{x}_Q\\ \mathbf{u}=\mathbf{u}_Q}}, \qquad \mathbf{B} = \left.\frac{\partial \mathbf{F}}{\partial \mathbf{u}}\right|_{\substack{\mathbf{x}=\mathbf{x}_Q\\ \mathbf{u}=\mathbf{u}_Q}}, \qquad \mathbf{C} = \left.\frac{\partial \mathbf{G}}{\partial \mathbf{x}}\right|_{\substack{\mathbf{x}=\mathbf{x}_Q\\ \mathbf{u}=\mathbf{u}_Q}}, \qquad \mathbf{D} = \left.\frac{\partial \mathbf{G}}{\partial \mathbf{u}}\right|_{\substack{\mathbf{x}=\mathbf{x}_Q\\ \mathbf{u}=\mathbf{u}_Q}} \tag{24.26}$$

The linearization ideas presented above are illustrated in the following example.

Example 24.2

Consider the magnetic levitation system shown in Fig. 24.2.

The metallic sphere is subject to two forces: its own weight, mg, and the attraction force generated by the electromagnet, $f(t)$. The electromagnet is commanded through a voltage source, $e(t) > 0$, $\forall t$.

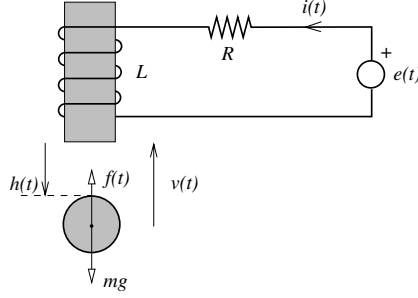



FIGURE 24.2 Magnetic levitation system.

The attraction force on the sphere, $f(t)$, depends on the distance $h(t)$ and the current, $i(t)$. This relation can be approximately described by

$$f(t) = \frac{K_1}{h(t) + K_2} i(t) \tag{24.27}$$

where K_1 and K_2 are positive constants.

Using first principles we can write

$$e(t) = Ri(t) + L\frac{di(t)}{dt} \tag{24.28}$$

$$v(t) = -\frac{dh(t)}{dt} \tag{24.29}$$

$$f(t) = \frac{K_1}{h(t) + K_2} i(t) = mg + m\frac{dv(t)}{dt} \tag{24.30}$$

We next choose as state variables: the current $i(t)$, the sphere position $h(t)$, and the sphere speed $v(t)$, i.e.,

$$x(t) = [x_1(t) \quad x_2(t) \quad x_3(t)]^T = [i(t) \quad h(t) \quad v(t)]^T \tag{24.31}$$

Then, from (24.28)–(24.30) we can set the system description as in (24.1) yielding

$$\frac{di(t)}{dt} = \frac{dx_1(t)}{dt} = -\frac{R}{L}x_1(t) + \frac{1}{L}e(t) \tag{24.32}$$

$$\frac{dh(t)}{dt} = \frac{dx_2(t)}{dt} = -x_3(t) \tag{24.33}$$

$$\frac{dv(t)}{dt} = \frac{dx_3(t)}{dt} = \frac{K_1}{m(x_2(t) + K_2)}x_1(t) - g \tag{24.34}$$

Before one can build the linearized model, an equilibrium point has to be computed. The driving input in this system is the source voltage $e(t)$. Say that the equilibrium point is obtained with $e(t) = E_Q$.

Hence, the state in equilibrium can be computed from (24.32) to (24.34), setting all the derivatives equal to zero, i.e.,

$$-\frac{R}{L}x_{1Q} + \frac{1}{L}E_Q = 0 \Rightarrow x_{1Q} = \frac{E_Q}{R} \tag{24.35}$$

$$-x_{3Q} = 0 \Rightarrow x_{3Q} = 0 \tag{24.36}$$

$$\frac{K_1}{m(x_{2Q}+K_2)}x_{1Q} - g = 0 \Rightarrow x_{2Q} = \frac{K_1}{mg}x_{1Q} - K_2 = \frac{K_1 E_Q}{mgR} - K_2 \tag{24.37}$$

The setting now is adequate to build the linearized model in the incremental input $\Delta e(t)$ and the incremental state $\Delta \mathbf{x}(t) = [\Delta x_1(t) \quad \Delta x_2(t) \quad \Delta x_3(t)]^T$. The result is

$$\frac{d\Delta x_1(t)}{dt} = -\frac{R}{L}\Delta x_1(t) + \frac{1}{L}\Delta e(t) \tag{24.38}$$

$$\frac{d\Delta x_2(t)}{dt} = -\Delta x_3(t) \tag{24.39}$$

$$\frac{d\Delta x_3(t)}{dt} = \frac{Rg}{E_Q}\Delta x_1(t) - \frac{Rmg^2}{K_1 E_Q}\Delta x_2(t) \tag{24.40}$$

If we define as the system output, the sphere position $h(t)$, we can then compare the above equations with (24.23) and (24.24) to obtain

$$\mathbf{A} = \begin{bmatrix} -\frac{R}{L} & 0 & 0 \\ 0 & 0 & -1 \\ \frac{Rg}{E_Q} & -\frac{Rmg^2}{K_1 E_Q} & 0 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} \frac{1}{L} \\ 0 \\ 0 \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \qquad \mathbf{D} = 0 \tag{24.41}$$

In the sequel we will drop the prefix Δ, but the reader should bear in mind that the model above is linear in the **incremental** components of the state, the inputs and the outputs around a chosen equilibrium point.

Linear State Space Models

Our starting point is now the linear time invariant state space model

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \tag{24.42}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \tag{24.43}$$

The solution to Eq. (24.42), subject to $\mathbf{x}(t_o) = \mathbf{x_o}$, is given by

$$\mathbf{x}(t) = e^{\mathbf{A}(t-t_o)}\mathbf{x_o} + \int_{t_o}^{t} e^{\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{u}(\tau)\,d\tau \quad \forall t \geq t_o \tag{24.44}$$

where the **transition matrix** $e^{\mathbf{A}t}$ satisfies

$$e^{\mathbf{A}t} = \mathbf{I} + \sum_{k=1}^{\infty} \frac{1}{k!} \mathbf{A}^k t^k \tag{24.45}$$

The interested reader can check that (24.44) satisfies (24.43). To do that he/she should use the Leibnitz's rule for the derivative of an integral.

With the above result, the solution for (24.43) is given by

$$\mathbf{y}(t) = \mathbf{C}e^{\mathbf{A}(t-t_o)}\mathbf{x_o} + \mathbf{C}\int_{t_o}^{t} e^{\mathbf{A}(t-\tau)}\mathbf{Bu}(\tau)\,d\tau + \mathbf{Du}(t) \tag{24.46}$$

System Dynamics

The state of the system has two components: the **unforced** component, $\mathbf{x_u}(t)$, and the **forced** component, $\mathbf{x_f}(t)$, where

$$\mathbf{x_u}(t) = e^{\mathbf{A}(t-t_o)}\mathbf{x_o} \tag{24.47}$$

$$\mathbf{x_f}(t) = \int_{t_o}^{t} e^{\mathbf{A}(t-\tau)}\mathbf{Bu}(\tau)\,d\tau \tag{24.48}$$

To gain insight into the state space model and its solution, consider the case when $t_o = 0$ and $u(t) = 0$ $\forall t \geq 0$, i.e., the state has only the **unforced part**. Then

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x_o} \tag{24.49}$$

Further assume that $\mathbf{A} \in \mathbb{R}^n$ and that, for simplicity, it has distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ with n (linearly independent) eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. Then there always exists a set of constants $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that

$$\mathbf{x_o} = \sum_{\ell=1}^{n} \alpha_\ell \mathbf{v}_\ell, \quad \alpha_\ell \in \mathbb{C} \tag{24.50}$$

A well-known result from linear algebra tells us that the eigenvalues of A^k are $\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k$ with corresponding eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. The application of this result yields

$$\mathbf{x}(t) = e^{\mathbf{A}t}\mathbf{x_o} = \mathbf{I} + \sum_{\ell=1}^{n} \alpha_\ell \sum_{k=1}^{\infty} \frac{1}{k!} \underbrace{\mathbf{A}^k \mathbf{v}_\ell}_{\lambda_\ell^k \mathbf{v}_\ell} t^k = \sum_{\ell=1}^{n} \alpha_\ell e^{\lambda_\ell t} \mathbf{v}_\ell \tag{24.51}$$

This equation shows that the unforced component of the state is a linear combination of **natural modes**, $\{e^{\lambda_\ell t}\}$, each of which is associated with an eigenvalue of **A**. Hence the matrix **A** determines:

- the structure of the unforced response
- the stability (or otherwise) of the system
- the speed of response

When the matrix **A** does not have a set of n independent eigenvectors, Jordan forms can be used (see, e.g., [9,10]).

Structure of the Unforced Response

In the absence of input, the state evolves as a combination of natural modes which belong to a defined class of functions: all those generated by exponentials with either real or complex exponents. Hence these modes include constants, real exponentials, pure sine waves, exponentially modulated sine waves, and some other specials functions arising from repeated eigenvalues.

To illustrate these ideas and their physical interpretation consider the system in Example 24.1. For that system

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\frac{K}{M} & -\frac{D}{M} \end{bmatrix} \tag{24.52}$$

Hence, the system eigenvalues are solutions to the equation

$$\det(\lambda\mathbf{I} - \mathbf{A}) = \lambda^2 + \frac{D}{M}\lambda + \frac{K}{M} = 0 \tag{24.53}$$

i.e.,

$$\lambda_{1,2} = -\frac{D}{2M} \pm \sqrt{\frac{D^2}{4M^2} - \frac{K}{M}} \tag{24.54}$$

Hence, when the damping is zero ($D = 0$), the system eigenvalues are a couple of conjugate imaginary numbers, and the two natural (complex) modes combine to yield a sustained oscillation with angular frequency $\omega_o = \sqrt{K/M}$. This is in agreement with our physical intuition, since we expect a sustained oscillation to appear when the system has nonzero initial conditions even if the external force, $f(t)$, is zero.

When the system is slightly damped ($D^2 < 4KM$), the matrix eigenvalues are conjugate complex numbers, and the associated complex natural modes combine to yield an exponentially damped sine wave. This also agrees with intuition, since the energy initially stored in the mass and the spring will periodically go from the mass to the spring and vice versa but, at the end, it will completely dissipate, as heat, in the viscous friction.

Finally if the damping is high ($D^2 > 4KM$), the matrix eigenvalues are a couple of negative real numbers, and the natural modes are two decaying exponentials. The heavy damping will preclude oscillations and the initial energy will dissipate quickly.

The three different situation are illustrated in Fig. 24.3. For this simulation we have used three different values of the viscous friction constant D and

$$M = 2 \text{ kg}, \quad K = 0.1 \text{ N/m}, \quad d(0) = 0.3 \text{ m}, \quad v(0) = 0.05 \text{ m/s} \tag{24.55}$$

Note that, except when there is no friction ($D = 0$), the mass comes to rest asymptotically.

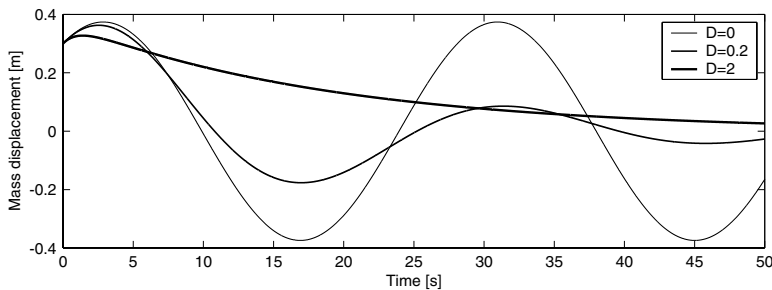



FIGURE 24.3 Unforced response of a mass-spring system.

Structure of the Forced Response

When the initial state is zero, the state will exhibit only the forced component. The forced component of the state will include natural modes and some additional **forced** or **particular** modes, which depend on the nature of the system input $\mathbf{u}(t)$. In general the forcing modes in the input will also appear in the state. However, some special cases arise when some of the forcing modes in $\mathbf{u}(t)$ coincide with some system natural modes.

System Stability

Stability in linear, time-invariant systems can also be analyzed using the state matrix **A**.

All systems variables can be expressed as linear functions of the state and the system input. When the system input $\mathbf{u}(t)$ is a vector of bounded time functions, then the boundeness of the system variables depends on the state to be bounded.

We then have the following result:

Theorem 24.1 *Consider a system with the state description* (24.42) *and* (24.43) *where* **B**, **C**, *and* **D** *have bounded elements. Then the system state (and hence the system output) is bounded for all bounded inputs if and only if the eigenvalues of* **A** *have negative real parts.*

To illustrate this theorem we again consider the magnetic levitation system from Example 24.2. For that system the matrix **A** (in the linearized model) is given by

$$\mathbf{A} = \begin{bmatrix} -\frac{R}{L} & 0 & 0 \\ 0 & 0 & -1 \\ \frac{Rg}{E_Q} & -\frac{Rmg^2}{K_1 E_Q} & 0 \end{bmatrix} \tag{24.56}$$

and its eigenvalues are the roots of $\det(\lambda\mathbf{I} - \mathbf{A}) = 0$, where

$$\det(\lambda\mathbf{I} - \mathbf{A}) = \left(\lambda + \frac{R}{L}\right)\left(\lambda - \sqrt{\frac{Rmg^2}{K_1 E_Q}}\right)\left(\lambda + \sqrt{\frac{Rmg^2}{K_1 E_Q}}\right) \tag{24.57}$$

One can then see that the set of matrix eigenvalues includes one which is **real and greater than zero**. This implies that the system is **unstable**. This is in agreement with physical reasoning. Indeed, at least theoretically, we can position the sphere in equilibrium (this is described by x_{2Q} in (24.37)). However, this is an unstable equilibrium point, since as soon as we slightly perturb the sphere, it accelerates either towards the ground or towards the magnet.

Speed of Response and Resonances

Even if the system is stable there are still some questions regarding other fundamental properties.

To start with, in stable systems the real part of the eigenvalues determines the speed at which the associated mode converges to zero. The slowest modes, the *dominant* modes, determine the speed at which the system output settles at its steady state value, i.e., determine the system speed of response. For example, if the system dominant eigenvalues are $\lambda_{1,2} = -\sigma \pm j\omega_o$, $\sigma > 0$, the combined natural modes generate an exponentially damped sine wave $y(t) = Ae^{-\sigma t}\sin(\omega_o t + \alpha)$. We then observe that this signal decays faster for a larger σ.

A second issue, of special importance for flexible structures, is the presence of resonances, which have associated complex eigenvalues. In physical systems, the existence of complex eigenvalues is intimately connected to the presence of two forms of energy. The resonance describes the (poorly damped) oscillation between those two forms of energy. In electric circuits those energies are the electrostatic energy in capacitors and the electromagnetic energy in inductors. In mechanical systems we have the kinetic energy of moving masses and the potential energy in springs. Flexible structures may have many resonant modes. One of the main problems with resonances occurs when the input contains energy at a frequency

close to the resonant frequency. For example, if a system has eigenvalues $\lambda_{1,2} = -0.05 \pm j$, i.e., the resonant frequency is 1 rad/s and, additionally, one of the **input** components is a sine wave of frequency 0.9 rad/s, then the system output exhibits a very large (forced) oscillation with amplitude initially growing almost linearly and later, stabilizing to a constant value. In real situations this phenomenon may destroy the system (recall the Tacoma bridge case).

State Similarity Transformation

We have already said that the choice of state variables is nonunique. Say that we have a system with input $\mathbf{u}(t)$, output $\mathbf{y}(t)$, and two different choices of state vectors: $\mathbf{x}(t) \in \mathbb{R}^n$ with an associated 4-tuple ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$), and $\bar{\mathbf{x}}(t) \in \mathbb{R}^n$ with an associated 4-tuple ($\bar{\mathbf{A}}$, $\bar{\mathbf{B}}$, $\bar{\mathbf{C}}$, $\bar{\mathbf{D}}$). Then there exists a nonsingular matrix $\mathbf{T} \in \mathbb{R}^{n \times n}$ such that

$$\bar{\mathbf{x}}(t) = \mathbf{T}\mathbf{x}(t) \Leftrightarrow \mathbf{x}(t) = \mathbf{T}^{-1}\bar{\mathbf{x}}(t) \tag{24.58}$$

This leads to the following equivalences:

$$\bar{\mathbf{A}} = \mathbf{T}\mathbf{A}\mathbf{T}^{-1}, \quad \bar{\mathbf{B}} = \mathbf{T}\mathbf{B}, \quad \bar{\mathbf{C}} = \mathbf{C}\mathbf{T}^{-1} \tag{24.59}$$

Different choices of state variables may or may not respond to different phenomenological approaches to the system analysis. Sometimes it is just a question of mathematical simplicity, as we shall see in section 24.6. In other occasions, the decision is made considering relative facility to measure certain system variables. However, what is important is that, no matter which state description is chosen, certain fundamental system characteristics do not change. They are related to the fact that the system eigenvalues are invariant with respect to similarity transformations, since

$$\det(\lambda\mathbf{I} - \bar{\mathbf{A}}) = \det(\lambda\mathbf{T}\mathbf{T}^{-1} - \mathbf{T}\mathbf{A}\mathbf{T}^{-1}) = \det(\mathbf{T})\det(\lambda\mathbf{I} - \mathbf{A})\det(\mathbf{T}^{-1}) \tag{24.60}$$

$$= \det(\lambda\mathbf{I} - \mathbf{A}) \tag{24.61}$$

Hence, stability, nature of the unforced response, and speed of response are invariants with respect to similarity transformations.

Example 24.3

Consider the electric network shown in Fig. 24.4

We choose the state vector $\mathbf{x}(t) = [x_1(t) \quad x_2(t)]^T = [i_L(t) \quad v_c(t)]^T$. Also $u(t) = v_f(t)$. Using first principles we have that

$$\frac{d\mathbf{x}(t)}{dt} = \begin{bmatrix} 0 & \frac{1}{L} \\ -\frac{1}{C} & -\frac{R_1 + R_2}{R_1 R_2 C} \end{bmatrix} \mathbf{x}(t) + \begin{bmatrix} 0 \\ \frac{1}{R_1 C} \end{bmatrix} u(t) \tag{24.62}$$

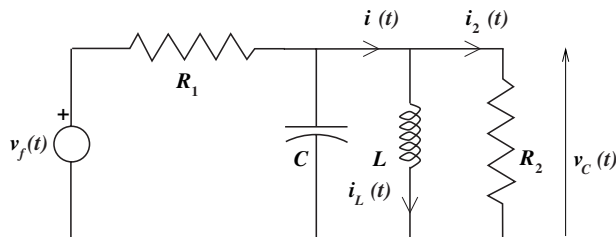



FIGURE 24.4 Electric network.

An alternative state vector is $\bar{\mathbf{x}}(t) = [\bar{x}_1(t) \quad \bar{x}_2(t)]^T = [i(t) \quad i_2(t)]^T$. It is straightforward to show that

$$\bar{\mathbf{x}}(t) = \underbrace{\frac{1}{R_2}\begin{bmatrix} R_2 & 1 \\ 0 & 1 \end{bmatrix}}_{\mathbf{T}} \mathbf{x}(t) \tag{24.63}$$

State Space and Transfer Functions

The state space description of linear time invariant systems is an alternative description to that provided by transfer functions. Strictly speaking, the state space description has a wider scope, as we shall see in this subsection.

For a linear time invariant system with input $\mathbf{u}(t) \in \mathbb{R}^m$ and output $\mathbf{y}(t) \in \mathbb{R}^p$, the transfer function, $\mathbf{H}(s) \in \mathbb{C}^{p \times m}$, is defined by the equation

$$\mathbf{Y}(s) = \mathbf{H}(s)\mathbf{U}(s), \quad \text{where} \quad [\mathbf{H}(s)]_{ij} = \frac{Y_i(s)}{U_j(s)} \tag{24.64}$$

i.e., the (i, j) element in matrix $\mathbf{H}(s)$ is the Laplace transformation of the response in the i^{th} output when a unit impulse is applied at the j^{th} input, with zero initial conditions and with the remaining inputs equal to zero for all $t \geq 0$.

On the other hand, if we Laplace-transform (24.42) and (24.43) with zero initial conditions, we obtain

$$\mathbf{X}(s) = (s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}\mathbf{U}(s) \tag{24.65}$$

$$\mathbf{Y}(s) = \mathbf{C}\mathbf{X}(s) + \mathbf{D}\mathbf{U}(s) = \underbrace{(\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D})}_{\mathbf{H}(s)}\mathbf{U}(s) \tag{24.66}$$

For simplicity, and to be able to go deeper into the analysis, **in the remaining part of this section we will focus our attention on the class of scalar systems**, i.e., systems with a single input and a single output (SISO systems). This means that $m = p = 1$, $\mathbf{B}$ becomes a column vector, $\mathbf{C}$ is a row vector, and $D = H(\infty)$ (in real systems it usually holds that $D = H(\infty) = 0$). For SISO systems, $H(s)$ is a quotient of polynomials in s, i.e.,

$$H(s) = \frac{\mathbf{C}\,\text{Adj}(s\mathbf{I} - \mathbf{A})\mathbf{B} + \mathbf{D}\,\det(s\mathbf{I} - \mathbf{A})}{\det(s\mathbf{I} - \mathbf{A})} \tag{24.67}$$

where Adj(o) denotes the adjoint matrix of (o).

A key issue is that the transfer function poles are eigenvalues of matrix $\mathbf{A}$. However, it is not true, in general, that the set of transfer function poles is identical to the set of matrix $\mathbf{A}$ eigenvalues. This can be appreciated through the following example.

Example 24.4

Let

$$\mathbf{A} = \begin{bmatrix} -2 & 1 \\ 0 & -3 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} 0 & 1 \end{bmatrix}, \qquad D = 0 \tag{24.68}$$

Then

$$H(s) = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} = \frac{1}{(s+2)(s+3)}\begin{bmatrix} 0 & 1 \end{bmatrix}\begin{bmatrix} s+3 & 1 \\ 0 & s+2 \end{bmatrix}\begin{bmatrix} 1 \\ 0.5 \end{bmatrix} \tag{24.69}$$

$$= \frac{0.5(s+2)}{(s+2)(s+3)} = \frac{0.5}{(s+3)} \tag{24.70}$$

Therefore, the transfer function has only one pole, although matrix **A** has two eigenvalues. We observe that there is a pole–zero cancellation in $H(s)$. This phenomenon is closely connected to the question of system properties, which is the central topic in section 24.6.

To acquire a phenomenological feeling on this issue, consider again the magnetic levitation system in Example 24.2. If we define the current $i(t)$ as the system output we can immediately see that the transfer function from the input $e(t)$ to this output has only one pole. This contrasts with the fact that the dimension of the state is equal to three. The explanation for this is that, in our simplified physical model, the current $i(t)$ is unaffected by the position and the speed of the metallic sphere (note that we have neglected the changes in the inductance due to changes in the sphere position).

The key result is that **the transfer function may not provide the same amount of information than the state space model** for the same system.

An interesting problem is to obtain a state space description from a given transfer function. The reader must be aware that the resulting state space model does not reveal pole-zero cancellations; for that reason, the obtained description is known as a **minimal realization.**

There are many methods to go from the transfer function to a state space model. We present below one of those methods.

Consider a transfer function given by

$$H_T(s) = \frac{B_o(s)}{A_o(s)} + H_T(\infty) = \frac{b_{n-1}s^{n-1} + b_{n-2}s^{n-2} + \cdots + b_1 s + b_0}{s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0} + H_T(\infty) \tag{24.71}$$

We first recall that $D = H_T(\infty)$. We can thus concentrate on the transfer function $H(s) = H_T(s) - H_T(\infty)$, which is a strictly proper transfer function.

Consider next a variable $v_\ell(t) \in \mathbb{R}$ whose Laplace transform, $V_\ell(s)$, satisfies

$$V_\ell(s) = \frac{s^{\ell-1}}{A_o(s)}U(s), \quad \ell \in \{1, 2, \ldots, n\} \tag{24.72}$$

This implies that

$$v_\ell(t) = \frac{dv_{\ell-1}(t)}{dt}, \quad \ell \in \{2, \ldots, n\} \tag{24.73}$$

$$Y(s) = \sum_{\ell=1}^{n} b_{\ell-1} V_\ell(s) \tag{24.74}$$

$$U(s) = \frac{A_o(s)}{A_o(s)}U(s) = \underbrace{\frac{s^n}{A_o(s)}U(s)}_{sV_n(s)} + \sum_{\ell=1}^{n} a_{\ell-1} \underbrace{\frac{s^{\ell-1}}{A_0(s)}U(s)}_{V_\ell(s)} \tag{24.75}$$

Now choose as state variables,

$$x_\ell(t) = v_\ell(t) \tag{24.76}$$

The above equations yield

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ -a_0 & -a_1 & -a_2 & \cdots & -a_{n-2} & -a_{n-1} \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \tag{24.77}$$

$$\mathbf{C} = [b_0 \quad b_1 \quad b_2 \quad \cdots \quad b_{n-1}], \qquad \mathbf{D} = H_T(\infty) \tag{24.78}$$

Example 24.5

The transfer function of a system is given by

$$H(s) = \frac{4s-10}{(s+2)^2(s-1)} = \frac{4s-10}{s^3+3s^2-4} \tag{24.79}$$

Then a minimal realization for this system is

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 4 & 0 & -3 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \tag{24.80}$$

$$\mathbf{C} = \begin{bmatrix} -10 & 4 & 0 \end{bmatrix}, \qquad \mathbf{D} = 0 \tag{24.81}$$

A key result is that a **system transfer function is invariant with respect to state similarity transformations**.

24.4 State Space Description for Discrete-Time and Sampled Data Systems

In this section we will present an overview of the state space description for discrete time systems, mainly based on the results presented for the continuous time case.

Discrete time models may arise from two different sources:

- From a *pure* discrete-time system, usually nonlinear, whose variables are defined only at specific time instants t_k. Systems like that can be found in economic systems, stochastic process theory, etc.
- From a discretization of a continuous-time system. In this case, we are only concerned with the value of some system variables at specific time instants. These models are useful when digital systems, such as microcontrollers, computers, PLCs, or others, interact with continuous-time *real systems* such as mechanical structures, valves, tanks, analog circuits or a whole industrial process[1]. These are called **sampled data systems**.

In both cases our analysis will be focused on the class of **linear and time invariant** models.

[1] Through *digital-to-analog* and *analog-to-digital* converters (DAC and ADC, respectively).

Linearization of Discrete Time Systems

The discrete time equivalents to (24.3) and (24.4) are given by the nonlinear equations

$$\mathbf{x}[t+1] = \mathbf{F_d}(\mathbf{x}[t], \mathbf{u}[t]) \tag{24.82}$$

$$\mathbf{y}[t] = \mathbf{G_d}(\mathbf{x}[t], \mathbf{u}[t]) \tag{24.83}$$

The linearization of models for discrete time systems follows along the same lines to that for continuous ones. Consider firstly an equilibrium point given by $\{\mathbf{x_Q}, \mathbf{u_Q}, \mathbf{y_Q}\}$:

$$\mathbf{x_Q} = \mathbf{F}_d(\mathbf{x_Q}, \mathbf{u_Q}) \tag{24.84}$$

$$\mathbf{y_Q} = \mathbf{G}_d(\mathbf{x_Q}, \mathbf{u_Q}) \tag{24.85}$$

Note that an equilibrium point is defined by a set of **constant** values of the state and **constant** values of the input which satisfy (24.82) and (24.83). This yields a constant system output. The discrete model can then be linearized around this equilibrium point. Defining

$$\Delta\mathbf{x}[t] = \mathbf{x}[t] - \mathbf{x_Q}, \qquad \Delta\mathbf{u}[t] = \mathbf{u}[t] - \mathbf{u_Q}, \qquad \Delta\mathbf{y}[t] = \mathbf{y}[t] - \mathbf{y_Q} \tag{24.86}$$

we have the state space model

$$\Delta\mathbf{x}[t+1] = \mathbf{A_d}\Delta\mathbf{x}[t] + \mathbf{B_d}\Delta\mathbf{u}[t] \tag{24.87}$$

$$\Delta\mathbf{y}[t] = \mathbf{C_d}\Delta\mathbf{x}[t] + \mathbf{D_d}\Delta\mathbf{u}[t] \tag{24.88}$$

where

$$\mathbf{A_d} = \left.\frac{\partial \mathbf{F_d}}{\partial \mathbf{x}}\right|_{\substack{\mathbf{x}=\mathbf{x_Q}\\ \mathbf{u}=\mathbf{u_Q}}}, \qquad \mathbf{B_d} = \left.\frac{\partial \mathbf{F_d}}{\partial \mathbf{u}}\right|_{\substack{\mathbf{x}=\mathbf{x_Q}\\ \mathbf{u}=\mathbf{u_Q}}}, \qquad \mathbf{C_d} = \left.\frac{\partial \mathbf{G_d}}{\partial \mathbf{x}}\right|_{\substack{\mathbf{x}=\mathbf{x_Q}\\ \mathbf{u}=\mathbf{u_Q}}}, \qquad \mathbf{D_d} = \left.\frac{\partial \mathbf{G_d}}{\partial \mathbf{u}}\right|_{\substack{\mathbf{x}=\mathbf{x_Q}\\ \mathbf{u}=\mathbf{u_Q}}} \tag{24.89}$$

Sampled Data Systems

As we have already said, discrete time models are frequently obtained by sampling inputs and outputs in continuous-time systems. When a digital device is to be used to act upon a continuous-time system, the command signals need only to be defined at specific instants, and not at *all* time. However, to be able to act upon the continuous-time system, we need a continuous-time signal. This is usually built with a *zero order hold*, which generates a staircase signal. Also, when we want to digitally measure a system variable this is done at some specific time instants. This means that we must **sample** the output signals. Figure 24.5 illustrates these concepts. If we assume a periodic sampling, with period Δ, we are only interested in the signals at time $k\Delta$. In the sequel we will drop Δ from the arguments, using $\mathbf{u}(k\Delta) = \mathbf{u}[t]$ for the input, $\mathbf{y}(k\Delta) = \mathbf{y}[t]$ for the output, and $\mathbf{x}(k\Delta) = \mathbf{x}[t]$ for the system state.

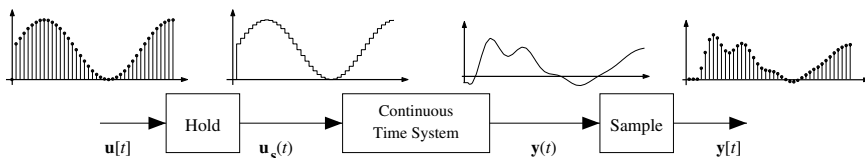



FIGURE 24.5 Schematic representation of a sampled data system.

If we consider the continuous, time-invariant, and linear state space model defined by equations (24.42) and (24.43), with initial state $\mathbf{x}(k_0\Delta) = \mathbf{x}_0$, we can use Eq. (24.44) to calculate the next value of the state:

$$\mathbf{x}(k_0\Delta + \Delta) = e^{\mathbf{A}(k_0\Delta+\Delta-k_0\Delta)}\mathbf{x}(k_0\Delta) + \int_{k_0\Delta}^{k_0\Delta+\Delta} e^{\mathbf{A}(k_0\Delta+\Delta-\tau)}\mathbf{B}\,\mathbf{u}(\tau)\,d\tau \tag{24.90}$$

Furthermore, if a zero order hold is used, i.e., $\mathbf{u}(\tau) = \mathbf{u}(k_0\Delta)$ for $k_0\Delta \le \tau < k_0\Delta + \Delta$, we obtain

$$\mathbf{x}(k_0\Delta + \Delta) = e^{\mathbf{A}\Delta}\mathbf{x}(k_0\Delta) + \int_0^{\Delta} e^{\mathbf{A}\eta}d\eta\mathbf{B}\,\mathbf{u}(k_0\Delta) \tag{24.91}$$

And, if we know the state and the input at time $k_0\Delta$, the output is defined by Eq. (24.43):

$$\mathbf{y}(k_0\Delta) = \mathbf{C}\mathbf{x}(k_0\Delta) + \mathbf{D}\,\mathbf{u}(k_0\Delta) \tag{24.92}$$

We can now conclude that given a continuous-time model with state space matrices $\{\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D}\}$, and we sample inputs and outputs every Δ seconds then, the equivalent sampled data systems will be described by the discrete-time state space model:

$$\mathbf{x}(k\Delta + \Delta) = \mathbf{A_d}\mathbf{x}(k\Delta) + \mathbf{B_d}\mathbf{u}(k\Delta) \tag{24.93}$$

$$\mathbf{y}(k\Delta) = \mathbf{C_d}\mathbf{x}(k\Delta) + \mathbf{D_d}\mathbf{u}(k\Delta) \tag{24.94}$$

where

$$\mathbf{A_d} = e^{\mathbf{A}\Delta}, \quad \mathbf{B_d} = \int_0^{\Delta} e^{\mathbf{A}\eta}d\eta\mathbf{B}, \quad \mathbf{C_d} = \mathbf{C}, \quad \mathbf{D_d} = \mathbf{D} \tag{24.95}$$

There are different methods to obtain $\mathbf{A_d}$ defined in (24.95), but a simple way to calculate this matrix is to use Laplace transformation. This yields

$$\mathbf{A_d} = e^{\mathbf{A}\Delta} = \mathcal{L}^{-1}\{(s\mathbf{I} - \mathbf{A})^{-1}\}\big|_{t=\Delta} \tag{24.96}$$

Example 24.6

Consider the mechanical system of Example 24.1 on the page 4, that was described by the state space model:

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\frac{K}{M} & -\frac{D}{M} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ -\frac{1}{M} \end{bmatrix} f(t) \tag{24.97}$$

where $f(t)$ is the external force, and where we can choose either the mass position, $x_1(t)$, or the mass velocity, $x_2(t)$, of the mass, as the system output.

For the purpose of a numerical illustration, we set $M = 1$ kg, $D = 1.2$ N s/m, and $K = 0.32$ N/m.

The matrix $\mathbf{A_d}$ is obtained from (24.96), applying inverse Laplace transformation

$$\mathbf{A_d} = \mathcal{L}^{-1}\left\{\begin{bmatrix} s & -1 \\ 0.32 & s+1.2 \end{bmatrix}^{-1}\right\}\Bigg|_{t=\Delta} = \begin{bmatrix} 2e^{-0.4\Delta} - e^{-0.8\Delta} & 2.5(e^{-0.4\Delta} - e^{-0.8\Delta}) \\ 0.8(e^{-0.4\Delta} - e^{-0.8\Delta}) & -e^{-0.4\Delta} + 2e^{-0.8\Delta} \end{bmatrix} \tag{24.98}$$

and the $\mathbf{B_d}$ matrix is obtained from (24.95):

$$\mathbf{B_d} = \int_0^{\Delta} \begin{bmatrix} 2e^{-0.4\eta} - e^{-0.8\eta} & 2.5(e^{-0.4\eta} - e^{-0.8\eta}) \\ 0.8(e^{-0.4\eta} - e^{-0.8\eta}) & -e^{-0.4\eta} + 2e^{-0.8\eta} \end{bmatrix} d\eta \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$
$$\Rightarrow \mathbf{B_d} = \begin{bmatrix} -6.25e^{-0.4\Delta} + 3.125e^{-0.8\Delta} + 3.125 \\ 2.5(e^{-0.4\Delta} - e^{-0.8\Delta}) \end{bmatrix} \tag{24.99}$$

Note that both, $\mathbf{A_d}$ and $\mathbf{B_d}$ are functions of Δ. Thus, the sampling period, Δ, has a strong presence in the dynamic behavior of the sampled system, as we shall observe in the following subsections.

Linear State Space Models

We will analyze the linear time invariant state space model

$$\mathbf{x}[t+1] = \mathbf{A_d}\mathbf{x}[t] + \mathbf{B_d}\mathbf{u}[t] \tag{24.100}$$

$$\mathbf{y}[t] = \mathbf{C_d}\mathbf{x}[t] + \mathbf{D_d}\mathbf{u}[t] \tag{24.101}$$

This can be a linearized discrete time model like (24.87) and (24.88), or a sampled data system like (24.93) and (24.94) where Δ has been dropped from the time argument.

The solution to Eqs. (24.100) and (24.101), subject to $\mathbf{x}[t_o] = \mathbf{x_o}$, is given by

$$\mathbf{x}[t] = \mathbf{A_d}^{(t-t_o)}\mathbf{x_o} + \sum_{i=0}^{(t-t_o)-1} \mathbf{A_d}^{(t-t_o)-i-1}\mathbf{B_d}\mathbf{u}[i+t_o] \quad \forall t \geq t_o \tag{24.102}$$

where $\mathbf{A_d}^{(t-t_o)}$ is the **transition matrix**.

The reader can check easily that (24.102) satisfies (24.100). With the above result, the solution for (24.101) is given by

$$\mathbf{y}[t] = \mathbf{C_d}\mathbf{A_d}^{(t-t_o)}\mathbf{x_o} + \mathbf{C_d}\sum_{i=0}^{(t-t_o)-1} (\mathbf{A_d}^{(t-t_o)-i-1}\mathbf{B_d}\mathbf{u}[i+t_o]) + \mathbf{D_d}\mathbf{u}[t] \tag{24.103}$$

System Dynamics

The state of the system has two components: the **unforced** component, $\mathbf{x_u}[t]$, and the **forced** component, $\mathbf{x_f}[t]$, where

$$\mathbf{x_u}[t] = \mathbf{A_d}^{(t-t_o)}\mathbf{x_o} \tag{24.104}$$

$$\mathbf{x_f}[t] = \sum_{\tau=0}^{(t-t_o)-1} \mathbf{A_d}^{(t-t_o)-i-1}\mathbf{B_d}\mathbf{u}[i+t_o] \tag{24.105}$$

To gain insight into the state space model and its solution consider the case when $t_o = 0$ and $u[t] = 0$, $\forall t \geq 0$, i.e., the state has only the **unforced part**. Then

$$\mathbf{x}[t] = \mathbf{A_d^t}\mathbf{x_o} \tag{24.106}$$

Further assume that $\mathbf{A_d} \in \mathbb{R}^{n \times n}$ and that, for simplicity, it has n distinct eigenvalues η_ℓ, with n linearly independent eigenvectors $\mathbf{v}_\ell$. Then there always exists a set of n constants α_ℓ such that

$$\mathbf{x_o} = \sum_{\ell=1}^{n} \alpha_\ell \mathbf{v}_\ell, \quad \alpha_\ell \in \mathbb{C} \tag{24.107}$$

A well known result from linear algebra tells us that the eigenvalues of $\mathbf{A_d^k}$ are η_ℓ^k, for $k \in \mathbb{N}$, with corresponding eigenvectors $\mathbf{v}_\ell$. The application of this result yields

$$\mathbf{x}[t] = \mathbf{A_d^t x_o} = \mathbf{A_d^t} \sum_{\ell=1}^{n} \alpha_\ell \mathbf{v}_\ell = \sum_{\ell=1}^{n} \alpha_\ell \underbrace{\mathbf{A_d^t V}_\ell}_{\eta_\ell^t \mathbf{v}_\ell} \tag{24.108}$$

$$\mathbf{x}[t] = \sum_{\ell=1}^{n} \alpha_\ell \eta_\ell^t \mathbf{v}_\ell \tag{24.109}$$

This equation shows that the unforced component of the state is a linear combination of **natural modes**, $\{\eta_\ell^t\}$, and each one is associated with an eigenvalue of $\mathbf{A_d}$, which are also known as **natural frequencies** of the model. Thus, we again have that the matrix $\mathbf{A_d}$ determines:

- the structure of the unforced response
- the stability (or otherwise) of the system
- the speed of response

Structure of the Unforced Response

In the absence of input, the state evolves as a combination of natural modes which belong to a defined class of functions: the powers of the model eigenvalues, either real or complex. These modes are discrete functions related to constants, real exponentials, pure sine waves, exponentially modulated sine waves, and some other specials functions arising from repeated eigenvalues.

To illustrate these ideas and their physical interpretation consider the sampled system in Example 24.6. If $\Delta = 1$, the state space matrices are

$$\mathbf{A_d} = \begin{bmatrix} 0.8913 & 0.5525 \\ -0.1768 & 0.2283 \end{bmatrix}, \qquad \mathbf{B_d} = \begin{bmatrix} 0.3397 \\ 0.5525 \end{bmatrix} \tag{24.110}$$

Hence, the system eigenvalues are solutions to the equation

$$\det(\eta \mathbf{I} - \mathbf{A_d}) = \det\left(\begin{bmatrix} \eta - 0.8913 & -0.5525 \\ 0.1768 & \eta - 0.2283 \end{bmatrix}\right) \tag{24.111}$$

$$= (\eta - 0.6703)(\eta - 0.4493) = 0 \tag{24.112}$$

i.e., $\eta_1 = 0.6703$, $\eta_2 = 0.4493$, and the unforced response is

$$\mathbf{x_u}[t] = C_1(0.6702)^t + C_2(0.4493)^t \tag{24.113}$$

where $C_1$and C_2 depend on the initial conditions only. We can observe that, when t tends to infinity, $\mathbf{x_u}[t]$ decays to zero, because $|\eta_{1,2}| < 1$. Also these eigenvalues are positive real numbers, so there is no oscillation

in the natural modes. This last observation is consistent with the parameter choice in Example 1.6, which made the mass-spring system to be overdamped.

Structure of the Forced Response

Consider the Eq. (24.102). Then, when the initial state is zero, the state will only exhibit the forced component. However, the forced component will still include natural modes plus some additional **forced** or **particular** modes, which depend on the nature of the system input $\mathbf{u}[t]$. In general, the forcing modes in the input will also appear in the state. However, special cases arise when a forcing mode in $\mathbf{u}[t]$ coincides with a system natural mode.

System Stability

Stability in linear time-invariant systems can also be analyzed using the state matrix $\mathbf{A_d}$. As we said, all systems variables can be expressed as linear functions of the state and the system input. When the system input $\mathbf{u}[t]$ is a vector of bounded time functions, then the boundedness of the system variables depends on the state to be bounded. We then have the following result:

Theorem 24.2 *Consider a system with the state description* (24.100) *and* (24.101) *where* $\mathbf{B_d}$, $\mathbf{C_d}$, *and* $\mathbf{D_d}$ *have bounded elements. Then the system state is bounded for all bounded inputs if and only if the eigenvalues of* $\mathbf{A_d}$ *lies inside the unit disc, i.e.,* $|\eta_\ell| < 1, \forall \ell$.

Speed of Response and Resonances

We recall that the natural modes of discrete-time systems are the powers of the eigenvalues η_ℓ. Since those eigenvalues can always be described as complex quantities, we can then write the natural modes as

$$(\eta_\ell)^t = (|\eta_\ell| e^{j\theta_\ell})^t = |\eta_\ell|^t e^{j\theta_\ell t}, \quad \text{where } \theta_\ell = \angle \eta_\ell \tag{24.114}$$

Therefore, we have that

- $0 < |\eta_\ell| < \infty$ determines the *speed* at which the mode decays to zero for stable systems ($|\eta_\ell| < 1$), or grows to infinity for unstable systems ($|\eta_\ell| > 1$)
- $-\pi < \theta_\ell \le \pi$ determines the *frequency* of the natural mode, measured in radians.

Although the natural modes of stable systems decay to zero, their nature determines the system transient response.

To illustrate these issues the **step response**, with zero initial conditions, is frequently used.

Example 24.7

Consider the first order, single-input single-output discrete-time system

$$\mathbf{x}[t+1] = \eta_\ell \mathbf{x}[t] + \mathbf{u}[t] \tag{24.115}$$

$$\mathbf{y}[t] = (1-\eta_\ell)\mathbf{x}[t] \tag{24.116}$$

To obtain the step response, we can use the Eq. (24.103), where $\mathbf{x_o} = 0$, $\mathbf{u}[t] = 1, \forall t \ge 0$.

$$\mathbf{y}[t] = \mathbf{C_d}\left(\sum_{i=0}^{t-1} \mathbf{A_d}^{t-i-1}\right)\mathbf{B_d} \tag{24.117}$$

$$= (1-\eta_\ell)\left(\sum_{i=0}^{t-1} \eta_\ell^{t-i-1}\right) = (1-\eta_\ell)\eta_\ell^{t-1}\frac{1-\eta_\ell^{-t}}{1-\eta_\ell^{-1}} \tag{24.118}$$

$$= 1-\eta_\ell^t \tag{24.119}$$

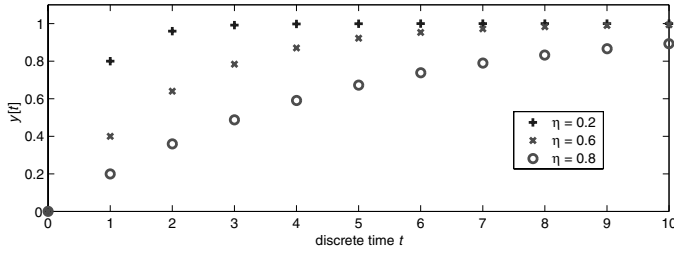



FIGURE 24.6 Step response of the system for different eigenvalues.

The output signal, $y[t] = y_h[t] + y_p[t]$, is shown in Fig. 24.6, for different values of the eigenvalue η_ℓ. The transient is given by $\mathbf{y_h}[t] = -\eta_\ell^t$, and the steady state response by $\mathbf{y_p}[t] = 1$.

We observed in Eq. (24.114) that the system eigenvalues define the damping of its transient response, but also determine its frequency of oscillation (when the eigenvalues have a nonzero imaginary part). The potential problem when resonant modes exist is the same problem we found in the context of continuous-time systems, i.e., the system input contains a sine wave or another kind of signal, with energy at a frequency close to one of the natural frequencies of the system. The system output still remains bounded, although it grows to undesirable amplitudes.

Example 24.8

Consider the discrete-time system described by the state space model

$$\mathbf{x}[t+1] = \begin{bmatrix} 1.2796 & -0.81873 \\ 1 & 0 \end{bmatrix} \mathbf{x}[t] + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u[t] \tag{24.120}$$

$$y[t] = \begin{bmatrix} 0 & 0.5391 \end{bmatrix} \mathbf{x}[t] \tag{24.121}$$

The eigenvalues of the system are obtained from $\mathbf{A_d}$:

$$\eta_{1,2} = 0.6398 \pm j0.6398 = 0.9048 \; (e^{j\pi/4}) \tag{24.122}$$

And the associated natural modes, present in the transient response, are

$$\eta_{1,2}^t = 0.9048^t e^{j\frac{\pi}{4}t} = 0.9048^t \left[\cos\left(\frac{\pi}{4}t\right) \pm j\sin\left(\frac{\pi}{4}t\right) \right] \tag{24.123}$$

The natural modes are slightly damped, because $|\eta_{1,2}|$ is close to 1, and they show an oscillation of frequency $\pi/4$.

In the plots shown in Fig. 24.7 we appreciate a strongly resonant output. The upper plot corresponds to an input $u[t] = \sin(\frac{\pi}{4}t)$, i.e., the input frequency coincides with the frequency of the natural modes. In the lower plot the input is a square wave of frequency input signal $\pi/12$. In this case, the input **third** harmonic has a frequency equal to the frequency of the natural modes.

Effect of Different Sampling Periods

We observe in Eq. (24.95) that $\mathbf{A_d}$ and $\mathbf{B_d}$ depend on the choice of the sampling period Δ. This choice determines the position of the eigenvalues of the system too. If we look at the Eq. (24.96), assuming that $\mathbf{A}$ has been diagonalized, we have that

$$\mathbf{A_d} = e^{\text{diag}\{\lambda_1,\ldots,\lambda_n\}\Delta} = \text{diag}\{e^{\lambda_1\Delta},\ldots, e^{\lambda_n\Delta}\} \tag{24.124}$$

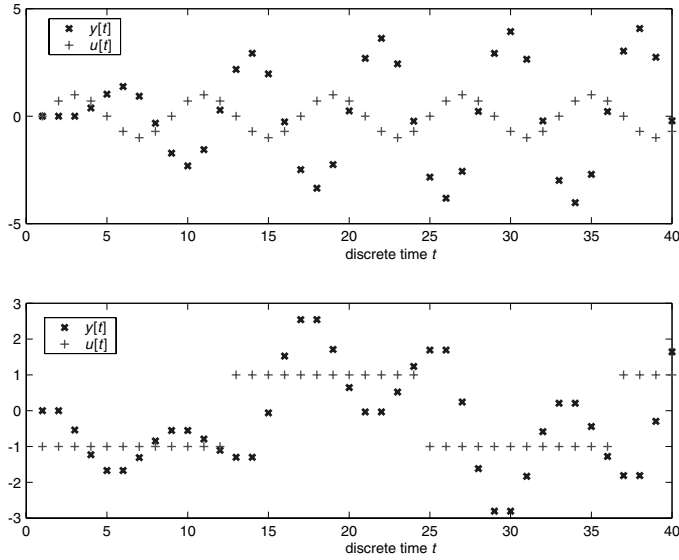



FIGURE 24.7 Resonant effect in the system output.

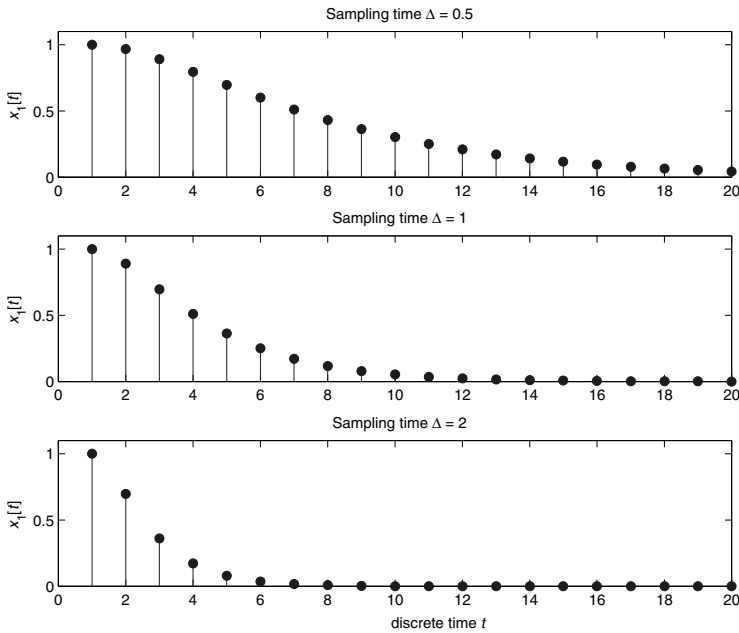



FIGURE 24.8 Effect of sampling in natural modes.

where $\{\lambda_1,\ldots, \lambda_n\}$are the eigenvalues of the underlying continuous-time systems. Then, these eigenvalues are mapped to the eigenvalues of the sampled-data system by equation:

$$\eta_\ell = e^{\lambda_\ell \Delta} \tag{24.125}$$

In Fig. 24.8 we observe the response of the sampled system of Example 24.6, choosing $x_1[t]$ as the system output, when the initial condition is $\mathbf{x}_o = [1 \quad 0]^T$, for different values of Δ. Observe that the horizontal axis corresponds to t, so the *real* instants times are $t\Delta$.

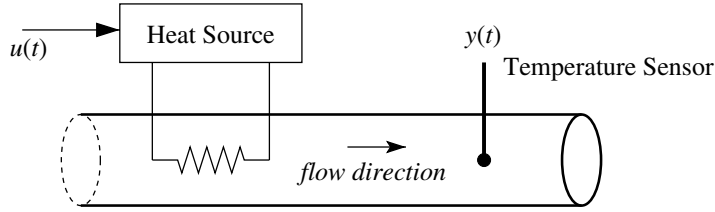



FIGURE 24.9 Heating system with time delay.

A fundamental issue regarding sampling of continuous-time signals is that the sampling period has to be chosen small enough to capture the essential nature of the signal to be sampled. To exemplify an ill-chosen Δ, assume that the signal $f(t) = A\sin(\omega_o t)$ is sampled every Δ seconds, with $\Delta = 2\ell\pi/\omega_o$, $\ell \in \mathbb{N}$. Then the resulting discrete time signal is $f[t] = 0$, $\forall t \in \mathbb{Z}$.

Sampled Data Systems and Time Delays

We said in section 24.3 that one cannot use continuous-time state space models to describe systems with time delays, because they are infinite dimensional systems. It was also said there that we would be able to tackle this problem using sampled signals. This is done using the following example.

Example 24.9

Consider the heating system sketched in Fig. 24.9.

The measured temperature, $y(t)$, of the flow depends on the power injected by the heat source. This source is commanded by a control signal $u(t)$. Changes in $u(t)$ yield changes in the temperature $y(t)$, but with a significant time delay. The linearized system can thus be represented by the transfer function:

$$\frac{Y(s)}{U(s)} = H(s) = \frac{e^{-\tau s}K}{s+\lambda} \tag{24.126}$$

where $U(s)$ and $Y(s)$ are the Laplace transforms of $u(t)$ and $y(t)$, respectively.

We next assume that the input and output signals are sampled every $\Delta[s]$. The time delay τ, in seconds, is a function of the flow velocity and we can assume, for simplicity, that τ it is a multiple of the sampling interval Δ, i.e., $\tau = m\Delta, m \in \mathbb{Z}^+$. These delays translate in a factor z^m in the denominator of the Z-transform transfer function. In other words, the delay gives rise to a set of m poles at the origin. Furthermore, the continuous-time system eigenvalue at $s = -\lambda$ becomes a discrete-time system eigenvalue at $z = e^{-\lambda\Delta}$ (see Eq. (24.125)). The resulting transfer function is

$$\frac{Y[z]}{U[z]} = H[z] = \frac{K}{\lambda}\frac{1-e^{-\lambda\Delta}}{z^m(z-e^{-\lambda\Delta})} \tag{24.127}$$

And this transfer function can be expressed as the discrete state space model

$$x_1[t+1] = x_2[t] \tag{24.128}$$

$$x_2[t+1] = x_3[t] \tag{24.129}$$

$$\vdots \qquad \vdots$$

$$x_m[t+1] = x_{m+1}[t] \tag{24.130}$$

$$x_{m+1}[t+1] = e^{-\lambda\Delta}x_{m+1}[t] + \frac{K}{\lambda}(1-e^{-\lambda\Delta})u[t] \tag{24.131}$$

$$y[t] = x_1[t] \tag{24.132}$$

We can then think of the states variables $x_{m+1}[t], \ldots, x_1[t]$ as the temperature at equally spaced points, between the heat source and the temperature sensor.

When the time delay τ is not a multiple of the sampling period Δ, an additional pole at the origin and an additional zero appear in the discrete transfer function. The details can be found elsewhere, e.g., in [7].

State Similarity Transformation

The idea of transforming the state via a similarity transformation equally applies to discrete-time systems. The system properties also remain unchanged.

State Space and Transfer Functions

For discrete-time systems the relation between state space and transfer function models is basically the same as in the continuous-time case (see section "State Space and Transfer Functions"). As we said then, the state space description of linear time invariant systems is an alternative description to that provided by transfer functions, although in some situations it provides more information on the system.

For a linear discrete-time invariant system with input $\mathbf{u}[t] \in \mathbb{R}^m$ and output $\mathbf{y}[t] \in \mathbb{R}^p$, the transfer function, $\mathbf{H}[z] \in \mathbb{C}^{p \times m}$, is defined by the equation

$$\mathbf{Y}[z] = \mathbf{H}[z]\mathbf{U}[z], \quad \text{where} \quad [\mathbf{H}[z]]_{ij} = \frac{Y_i[z]}{U_j[z]} \tag{24.133}$$

i.e., the (i, j) element in matrix $\mathbf{H}[z]$ is the Zeta transformation of the response in the i^{th} output when a unit Kronecker's delta is applied at the j^{th} input, with zero initial conditions and with the remaining inputs equal to zero for all $t \geq 0$.

On the other hand, if we apply Zeta transform to the discrete time state space model (24.100) and (24.101), with zero initial conditions, we have

$$\mathbf{X}[z] = (z\mathbf{I} - \mathbf{A_d})^{-1}\mathbf{B_d}\mathbf{U}[z] \tag{24.134}$$

$$\mathbf{Y}[z] = \mathbf{C_d}\mathbf{X}[z] + \mathbf{D_d}\mathbf{U}[z] \tag{24.135}$$

leading to

$$\mathbf{C_d}(z\mathbf{I} - \mathbf{A_d})^{-1}\mathbf{B_d} + \mathbf{D_d} = \mathbf{H}[z] \tag{24.136}$$

In the following analysis, we will focus on the class of scalar systems, i.e., $m = p = 1$, $\mathbf{B_d}$, $\mathbf{C_d}^T$ are column vectors, and $\mathbf{D_d} = H[\infty]$. We can then see that $H[z]$ is a quotient of polynomials in z, i.e.,

$$H[z] = \frac{\mathbf{C_d}\,\mathrm{Adj}(z\mathbf{I} - \mathbf{A_d})\mathbf{B_d} + \mathbf{D_d}\det(z\mathbf{I} - \mathbf{A_d})}{\det(z\mathbf{I} - \mathbf{A_d})} \tag{24.137}$$

where Adj(o) denotes the adjoint matrix of (o).

We have again, paralleling the continuous-time case, that the transfer function poles are eigenvalues of $\mathbf{A_d}$. However, it is not true in general that the set of transfer function poles is identical to the set of eigenvalues of the matrix. It is important to realize that transfer function models can hide cancellations between poles and zeros, with the consequences described in subsections "Controllability, Reachability and Stabilizability" and "Observability, Reconstructability and Detectability."

A key result for discrete-time system is the same for continuous-time systems: **the transfer function may not provide the same amount of information than the state space model** for the same system.

One way to obtain the state space model is to use the same method proposed in section "State Space and Transfer Functions," applying Zeta transformation instead of Laplace transformation, and using the fact that

$$F[z] = \mathcal{Z}\{f[t]\} \Leftrightarrow zF[z] = \mathcal{Z}\{f[t+1]\} \tag{24.138}$$

Example 24.10

The transfer function of a system is given by

$$H[z] = \frac{2z^2 - z + 1}{(z-0.8)(z-0.6)} = \frac{1.8z + 0.04}{z^2 - 1.4z + 0.48} + 2 \tag{24.139}$$

Then a minimal realization for this system is

$$\mathbf{A_d} = \begin{bmatrix} 0 & 1 \\ -0.48 & -1.4 \end{bmatrix}, \qquad \mathbf{B_d} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \tag{24.140}$$

$$\mathbf{C_d} = [0.04 \quad 1.8], \qquad \mathbf{D_d} = 2 \tag{24.141}$$

In discrete-time models it also happens that the **system transfer function is invariant with respect to state similarity transformations**.

24.5 State Space Models for Interconnected Systems

To build state space models for complex systems it is sometimes useful (and possible) to describe them as the interconnection of simpler systems. That interconnection is usually a combination of three basic interconnection structures: series, parallel, and feedback. In those three basic cases our aim is to obtain a state space model for the composite system.

In the following analysis we will use two systems, which are defined by

$$\text{System 1:} \quad \frac{d\mathbf{x}_1(t)}{dt} = \mathbf{A}_1\mathbf{x}_1(t) + \mathbf{B}_1\mathbf{u}_1(t) \tag{24.142}$$

$$\mathbf{y}_1(t) = \mathbf{C}_1\mathbf{x}_1(t) + \mathbf{D}_1\mathbf{u}_1(t) \tag{24.143}$$

$$\text{System 2:} \quad \frac{d\mathbf{x}_2(t)}{dt} = \mathbf{A}_2\mathbf{x}_2(t) + \mathbf{B}_2\mathbf{u}_2(t) \tag{24.144}$$

$$\mathbf{y}_2(t) = \mathbf{C}_2\mathbf{x}_2(t) + \mathbf{D}_2\mathbf{u}_2(t) \tag{24.145}$$

Series Connection

The system interconnection shown in Fig. 24.10 is known as a series or cascade connection. To build the desired state space model, we first observe that $\mathbf{y}_2(t) = \mathbf{u}_1(t)$. Also, the composite system input is $\mathbf{u}(t) = \mathbf{u}_2(t)$,

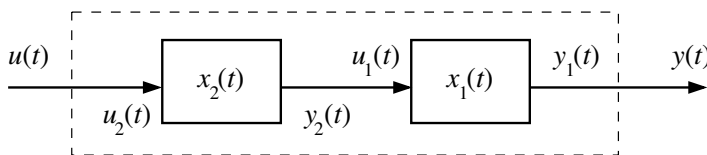



FIGURE 24.10 Series connection.

and the composite system output is $\mathbf{y}(t) = \mathbf{y}_1(t)$. We thus obtain

$$\begin{bmatrix} \dot{\mathbf{x}}_1(t) \\ \dot{\mathbf{x}}_2(t) \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{B}_1\mathbf{C}_2 \\ 0 & \mathbf{A}_2 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1(t) \\ \mathbf{x}_2(t) \end{bmatrix} + \begin{bmatrix} \mathbf{B}_1\mathbf{D}_2 \\ \mathbf{B}_2 \end{bmatrix} \mathbf{u}(t) \tag{24.146}$$

$$\mathbf{y}(t) = [\mathbf{C}_1 \quad \mathbf{D}_1\mathbf{C}_2] \begin{bmatrix} \mathbf{x}_1(t) \\ \mathbf{x}_2(t) \end{bmatrix} + [\mathbf{D}_1\mathbf{D}_2]\,\mathbf{u}(t) \tag{24.147}$$

Parallel Connection

The system interconnection shown in Fig. 24.11 is known as a parallel connection. To obtain the desired state space model we observe that the input is $\mathbf{u}(t) = \mathbf{u}_1(t) = \mathbf{u}_2(t)$ and the output for the whole system is $\mathbf{y}(t) = \mathbf{y}_1(t) + \mathbf{y}_2(t)$. We obtain

$$\begin{bmatrix} \dot{\mathbf{x}}_1(t) \\ \dot{\mathbf{x}}_2(t) \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1(t) \\ \mathbf{x}_2(t) \end{bmatrix} + \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} \mathbf{u}(t) \tag{24.148}$$

$$\mathbf{y}(t) = [\mathbf{C}_1 \quad \mathbf{C}_2] \begin{bmatrix} \mathbf{x}_1(t) \\ \mathbf{x}_2(t) \end{bmatrix} + [\mathbf{D}_1 + \mathbf{D}_2]\,\mathbf{u}(t) \tag{24.149}$$

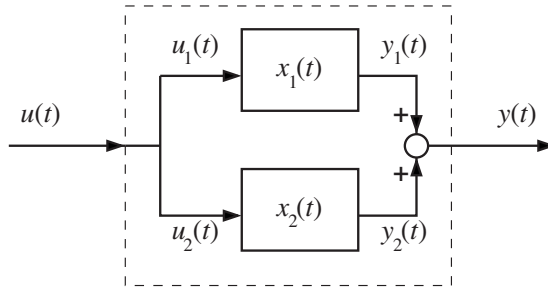



FIGURE 24.11 Parallel connection.

Feedback Connection

The system interconnection shown in Fig. 24.12 is known as feedback connection (with unit negative feedback), and it corresponds to the basic structure of a control loop, where S_1 is the *plant* and S_2 is the *controller*. To build the composite state space model we observe that the overall system input satisfies the equation $\mathbf{u}(t) = \mathbf{u}_2(t) + \mathbf{y}_1(t)$, and the overall system output is $\mathbf{y}(t) = \mathbf{y}_1(t)$. Furthermore, we assume

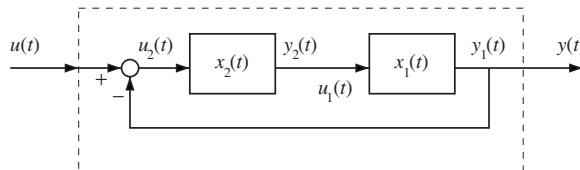



FIGURE 24.12 Feedback connection.

that the system S_1 (the plant) is strictly proper, i.e., $\mathbf{D}_1 = 0$. We then obtain

$$\begin{bmatrix} \dot{\mathbf{x}}_1(t) \\ \dot{\mathbf{x}}_2(t) \end{bmatrix} = \begin{bmatrix} \mathbf{A}_1 - \mathbf{B}_1\mathbf{D}_2\mathbf{C}_1 & \mathbf{B}_1\mathbf{C}_2 \\ -\mathbf{B}_2\mathbf{C}_1 & \mathbf{A}_2 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1(t) \\ \mathbf{x}_2(t) \end{bmatrix} + \begin{bmatrix} \mathbf{B}_1\mathbf{D}_2 \\ \mathbf{B}_2 \end{bmatrix} \mathbf{u}(t) \tag{24.150}$$

$$\mathbf{y}(t) = [\mathbf{C}_1 \quad \mathbf{0}] \begin{bmatrix} \mathbf{x}_1(t) \\ \mathbf{x}_2(t) \end{bmatrix} \tag{24.151}$$

The same results apply, mutatis mutandis, to discrete-time interconnected systems. More details can be found elsewhere, e.g., in [15].

24.6 System Properties

Controllability, Reachability, and Stabilizability

A very important question that we must be interested in regarding control systems using state space models is whether or not we can steer the state via the control input to certain locations in the state space. We must remember that the states of a system frequently are internal variables like temperature, pressure, level of tanks, or others. These are sometimes critical variables that we want to keep between specific values.

Controllability

The issue of controllability is concerned with whether or not a given initial state $\mathbf{x}_0$ can be steered to the origin in finite time using the input $\mathbf{u}(t)$.

Example 24.11

If we examine the model defined in (24.152), we note that the input $u(t)$ has no effect over the state $x_2(t)$.

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(t) \tag{24.152}$$

Given an initial state $[x_1(0), x_2(0)]^T$, the input $u(t)$ can be chosen to steer $x_1(t)$ to zero, while $x_2(t)$ remains unchanged.

Formally, we have the following definition:

Definition 24.1 A state x_o is said to **controllable** if there exists a finite interval $[0, T]$ and an input $\{u(t), t \in [0, T]\}$ such that $x(T) = 0$. If all states are controllable, then the system is said to be **completely controllable**.

Reachability

A related concept is that of **reachability**, used sometimes in discrete-time systems. It is formally defined as follows:

Definition 24.2 A state $\bar{\mathbf{x}} \neq 0$ is said to be **reachable**, from the origin, if given $\mathbf{x}(0) = 0$, there exists a finite time interval $[0, T]$ and an input $\{\mathbf{u}(t), t \in [0, T]\}$ such that $\mathbf{x}(T) = \bar{\mathbf{x}}$. If all states are reachable the system is said to be **completely reachable**.

For continuous, time-invariant, linear systems, there is no distinction between complete controllability and reachability. However, the following example illustrates that there is a subtle difference in the

discrete-time case. Consider the system and the output

$$\mathbf{x}[t+1] = \underbrace{\begin{bmatrix} 0.5 & 1 \\ -0.25 & -0.5 \end{bmatrix}}_{\mathbf{A_d}} \mathbf{x}[t] \Rightarrow \mathbf{x}[t] = \begin{bmatrix} 0.5 & 1 \\ -0.25 & -0.5 \end{bmatrix}^t \mathbf{x}[0] \tag{24.153}$$

We can see that this system is completely controllable since $\mathbf{x}[t] = 0$, $\forall t \geq 2$ and $\forall \mathbf{x}[0] \in \mathbb{R}^2$. This implies that every initial state is controllable. However, no nonzero state is reachable.

In view of the distinction between controllability and reachability in discrete time, we will use the term *controllability* in the sequel to cover the stronger of the two concepts.

Usually, in the context of linear time invariant systems, controllability and reachability are used interchangeably.

Controllability Test

We now present a systematic way to determine the complete controllability of a system.

Theorem 24.3 *Consider the linear, time-invariant, state space model where* $\mathbf{A} \in \mathbb{R}^{n \times n}$*:*

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \tag{24.154}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \tag{24.155}$$

i) *The set of all controllable states is the* ***range space*** *of the controllability matrix* $\mathbf{\Gamma}_c[\mathbf{A}, \mathbf{B}]$ *where*

$$\Gamma_c[\mathbf{A}, \mathbf{B}] \triangleq [\mathbf{B} \quad \mathbf{AB} \quad \mathbf{A}^2\mathbf{B} \quad \cdots \quad \mathbf{A}^{n-1}\mathbf{B}] \tag{24.156}$$

ii) *The model is completely controllable if and only if* $\mathbf{\Gamma}_c[\mathbf{A}, \mathbf{B}]$ *has* ***full row rank****.*

Example 24.12

Consider the state space model given in (24.152), with state space matrices

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{24.157}$$

The controllability matrix for this system, is given by

$$\Gamma_c[\mathbf{A}, \mathbf{B}] = [\mathbf{B} \quad \mathbf{AB}] = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \tag{24.158}$$

Clearly, rank $\mathbf{\Gamma}_c[\mathbf{A}, \mathbf{B}] = 1$, thus the system is **not** completely controllable.

The result above applies to continuous-time models, and it holds equally well for reachability of discrete-time models.

Also we can see that the controllability of a system is a property that does not depend on the choice of state variables. To see that, consider the similarity transformation defined in subsection "State Similarity Transformation." Then, observing that $\bar{\mathbf{A}}^i = \mathbf{T}^{-1}\mathbf{A}^i\mathbf{T}$, we have

$$\Gamma_c[\bar{\mathbf{A}}, \bar{\mathbf{B}}] = \mathbf{T}^{-1}\Gamma_c[\mathbf{A}, \mathbf{B}] \tag{24.159}$$

which implies that $\Gamma_c[\bar{\mathbf{A}}, \bar{\mathbf{B}}]$ and $\Gamma_c[\mathbf{A}, \mathbf{B}]$ have the same rank.

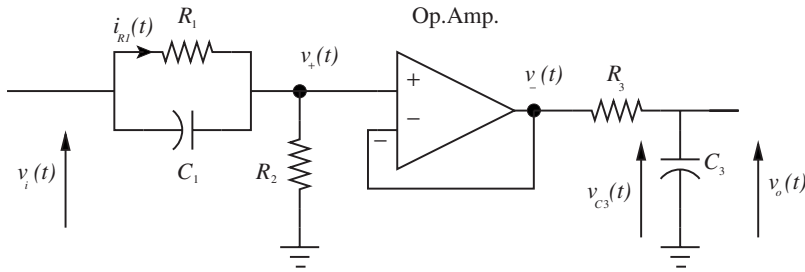



FIGURE 24.13 Electronic circuit.

The reader may wish to check that the state space models used to describe signals in subsection "Signals and State Space Description" are uncontrollable. Indeed, it is always true **that any state space model where B = 0 is completely uncontrollable**.

Loss of Controllability

Lack of controllability is sometimes a structural feature. However, in some other cases, it depends on the numerical value of certain parameters. We illustrate this in the following example.

Example 24.13

Consider the electronic circuit shown in Fig. 24.13.

We first build a state space model for the circuit. We choose, as state variables, $x_1(t) = i_{R1}(t)$ and $x_2(t) = v_{C3}(t)$. Using first principles on the left half of the circuit we have that

$$i_{C1} = C_1\frac{d}{dt}(v_i - v_+), \qquad i_{R1} = \frac{v_i - v_+}{R_1}, \qquad i_{R2} = \frac{v_+}{R_2}, \qquad i_{C1} = i_{R2} - i_{R1} \tag{24.160}$$

This yields

$$\frac{di_{R1}(t)}{dt} = -\frac{(R_1 + R_2)}{C_1R_1R_2}i_{R1}(t) + \frac{1}{C_1R_1R_2}v_i(t) \tag{24.161}$$

$$v_+(t) = -R_1 i_{R1}(t) + v_i(t) \tag{24.162}$$

And, similarly, from the right half of the circuit we obtain

$$\frac{dv_{C3}(t)}{dt} = -\frac{1}{R_3C_3}v_{C3}(t) + \frac{1}{R_3C_3}v_-(t) \tag{24.163}$$

$$v_o(t) = v_{C3}(t) \tag{24.164}$$

The (ideal) operational amplifier ensures that $v_+(t) = v_-(t)$, so we can combine the state space models given in Eqs. (24.161)–(24.164) to obtain

$$\begin{bmatrix} \frac{di_{R1}(t)}{dt} \\ \frac{dv_{C3}(t)}{dt} \end{bmatrix} = \begin{bmatrix} -\frac{(R_1+R_2)}{C_1R_1R_2} & 0 \\ -\frac{R_1}{R_3C_3} & -\frac{1}{R_3C_3} \end{bmatrix} \begin{bmatrix} i_{R1}(t) \\ v_{C3}(t) \end{bmatrix} + \begin{bmatrix} \frac{1}{C_1R_1R_2} \\ \frac{1}{C_3R_3} \end{bmatrix} v_i(t) \tag{24.165}$$

$$v_o(t) = [0 \quad 1]\begin{bmatrix} i_{R1}(t) \\ v_{C3}(t) \end{bmatrix} \tag{24.166}$$

The controllability matrix is then given by

$$\Gamma_c[\mathbf{A}, \mathbf{B}] = [\mathbf{B} \quad \mathbf{AB}] = \begin{bmatrix} \frac{1}{R_1 R_2 C_1} & \frac{-(R_1 + R_1)}{(R_1 R_2 C_1)^2} \\ \frac{1}{R_3 C_3} & \frac{-(R_2 C_1 + R_3 C_3)}{(R_3 C_3)^2 R_2 C_1} \end{bmatrix} \tag{24.167}$$

and

$$\det(\Gamma_c[\mathbf{A}, \mathbf{B}]) = \frac{R_2}{(R_1 R_2 R_3 C_1 C_2)^2}(-R_1 C_1 + R_3 C_3) \tag{24.168}$$

where we can observe that the system is completely controllable if, and only if, $R_1 C_1 \neq R_3 C_3$.

This issue has a very important interpretation if we analyze it from the transfer function point of view. Applying Laplace transform to Eqs. (24.161)–(24.164), the transfer function from $v_i(t)$ to $v_o(t)$ (recall that $V_+(s) = V_-(s)$) is given by

$$\frac{V_o(s)}{V_i(s)} = \frac{V_o(s)}{V_-(s)}\frac{V_+(s)}{V_i(s)} = \frac{\frac{1}{R_3 C_3}}{\left(s + \frac{1}{R_3 C_3}\right)} \cdot \frac{\left(s + \frac{1}{R_1 C_1}\right)}{\left(s + \frac{R_1 + R_2}{R_1 R_2 C_1}\right)} \tag{24.169}$$

where we can observe that the loss of complete controllability, when $R_1 C_1 = R_3 C_3$ obtained from (24.168), means that there is a **zero-pole cancellation** in the transfer function, i.e., the zero from the left half of the circuit in Fig. 24.13 is cancelled by the pole from the other part of the circuit. This issue will be discussed in more detail in section "Canonical Decomposition."

Controllability Gramian

The test of controllability gives us a *yes or no* answer about the controllability of a system model. However, to conclude that a system is completely controllable says nothing about the *degree* of controllability. For **stable** systems, we can quantify the effort to control the system state through the energy involved in the input signal $\mathbf{u}(t)$ applied from $t = -\infty$ to reach the state $\mathbf{x}(0) = \mathbf{x}_0$ at $t = 0$:

$$J(\mathbf{u}) = \int_{-\infty}^{0} \|\mathbf{u}(t)\|^2 dt = \int_{-\infty}^{0} \mathbf{u}(t)^T \mathbf{u}(t) dt \tag{24.170}$$

It can be shown that the minimal *control energy* is

$$J(\mathbf{u}_{\text{opt}}) = \mathbf{x}_o^T \mathbf{P}^{-1} \mathbf{x}_o \tag{24.171}$$

where

$$\mathbf{P} = \int_0^{\infty} e^{\mathbf{A}t} \mathbf{B} \mathbf{B}^T e^{\mathbf{A}^T t} dt \tag{24.172}$$

The matrix $\mathbf{P}$ is called the **controllability gramian**, and it measures the controllability of the state vector $\mathbf{x}(0)$. If this matrix is *small*, it means that we need a lot of energy in the control input $\mathbf{u}(t)$ to steer the state vector to $\mathbf{x}_0$. Indeed, we can appreciate the necessary effort for each one of the state variables, making, for example $\boldsymbol{x}_0 = [0,\ldots, 0, 1, 0,\ldots, 0]^T$.

It is important to emphasize that the existence of the integral defined in (24.172) is guaranteed only if the eigenvalues of **A** have negative real part, i.e., the system must be stable.

Also, the controllability gramian P defined in (24.172) satisfies the Lyapunov equation

$$\mathbf{AP} + \mathbf{PA}^T + \mathbf{BB}^T = 0 \tag{24.173}$$

For discrete-time systems we have the following equations for the controllability gramian:

$$\mathbf{P_d} = \sum_{k=0}^{\infty} \mathbf{A_d}^k \mathbf{B_d} \mathbf{B_d}^T (\mathbf{A_d}^T)^k \tag{24.174}$$

which satisfies

$$\mathbf{A_d P_d A_d}^T - \mathbf{P_d} + \mathbf{B_d B_d}^T = 0 \tag{24.175}$$

The sum defined in (24.174) is bounded if and only if the discrete-time system is stable, i.e., its eigenvalues lie inside the unit disc.

Example 24.14

We can analyze the model of the Example 24.13, where the electronic circuit was described by the state space models (24.165) and (24.166). If we want to appreciate the information that we can obtain from the controllability gramian, defined in (24.172), when the model is close to losing complete controllability, we can choose suitable values of the parameters that ensure $R_1C_1 \approx R_3C_3$.

If we choose

$$R_1 = R_2 = R_3 = 10^3\ \Omega, \qquad C_1 = 0.9 \times 10^3\ \mu\text{F}, \qquad C_3 = 10^3\ \mu\text{F} \tag{24.176}$$

the model will be described by

$$\begin{bmatrix} \dot{i}_{R1}(t) \\ \dot{v}_{C3}(t) \end{bmatrix} = \begin{bmatrix} -\frac{20}{9} & 0 \\ -10^3 & -1 \end{bmatrix} \begin{bmatrix} i_{R1}(t) \\ v_{C3}(t) \end{bmatrix} + \begin{bmatrix} \frac{0.01}{9} \\ 1 \end{bmatrix} v_i(t) \tag{24.177}$$

$$v_o(t) = [0 \quad 1] \begin{bmatrix} i_{R1}(t) \\ v_{C3}(t) \end{bmatrix} \tag{24.178}$$

If we look at the relative magnitude of the elements of **B**, we can *a priori* say that the effect of the input $u(t)$ upon the state $i_{R1}(t)$ will be much weaker than its effect upon the state $v_{C3}(t)$. To verify this we can compute the controllability gramian defined in (24.172), solving

$$\mathbf{0} = \mathbf{AP} + \mathbf{PA}^T + \mathbf{BB}^T \tag{24.179}$$

$$\mathbf{0} = \begin{bmatrix} -\frac{20}{9} & 0 \\ -10^3 & -1 \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} + \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \begin{bmatrix} -\frac{20}{9} & -10^3 \\ 0 & -1 \end{bmatrix} + \begin{bmatrix} \frac{0.01}{9} \\ 1 \end{bmatrix} \begin{bmatrix} \frac{0.01}{9} & 1 \end{bmatrix} \tag{24.180}$$

We have

$$\mathbf{P} = \begin{bmatrix} 0.28\times10^{-6} & 0.000258620 \\ 0.000258620 & 0.99999948 \end{bmatrix}, \qquad \mathbf{P}^{-1} = \begin{bmatrix} 4736624.0 & -1224.9 \\ -1224.9 & 1.3 \end{bmatrix} \tag{24.181}$$

So we can obtain the minimal control energy to steer the state $\mathbf{x}(t)$, from 0 in $t = -\infty$ to $\mathbf{x_0}$ in $t = 0$, from Eq. (24.171).

$$\mathbf{x_0} = [1, 0]^T \Rightarrow J(u_{\text{opt}}) = 4736624.0 \tag{24.182}$$

$$\mathbf{x_0} = [0, 1]^T \Rightarrow J(u_{\text{opt}}) = 1.3 \tag{24.183}$$

We can thus verify that the control energy to attain $i_{R1}(0) = 1$ is six orders of magnitude greater than the necessary energy to attain $v_{C3}(0) = 1$.

Also, if we substitute the parameter values in Eq. (24.169), we have that the transfer function is given by

$$\frac{V_o(s)}{V_i(s)} = \frac{1}{s+1} \cdot \frac{s+1+\frac{1}{9}}{s+\frac{20}{95}} \tag{24.184}$$

from where we observe a **zero-pole quasi cancellation**.

The idea of gramian has been extended to include the unstable case; see [16].

Canonical Decomposition and Stabilizability

If we have a system which is not completely controllable, it can be **decomposed** into a controllable subsystem and a completely uncontrollable subsystem in the following way.

Lemma 24.1 *Consider a system having rank* $\{\mathbf{\Gamma}_c[\mathbf{A}, \mathbf{B}]\} = k < n$. *Then there exists a similarity transformation* $\mathbf{T}$ *such that* $\bar{\mathbf{x}} = \mathbf{T}^{-1}\mathbf{x}$,

$$\bar{\mathbf{A}} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}, \qquad \bar{\mathbf{B}} = \mathbf{T}^{-1}\mathbf{B} \tag{24.185}$$

and $\bar{\mathbf{A}}$, $\bar{\mathbf{B}}$ *have the form*

$$\bar{\mathbf{A}} = \begin{bmatrix} \bar{\mathbf{A}}_c & \bar{\mathbf{A}}_{12} \\ 0 & \bar{\mathbf{A}}_{nc} \end{bmatrix}, \qquad \bar{\mathbf{B}} = \begin{bmatrix} \bar{\mathbf{B}}_c \\ 0 \end{bmatrix} \tag{24.186}$$

where $\bar{\mathbf{A}}_c$ *has dimension k and* $(\bar{\mathbf{A}}_c, \bar{\mathbf{B}}_c)$ *is completely controllable.*

The above result tells us what states we can and what states we cannot steer to zero. To appreciate this, we express the state and output equations in the form

$$\begin{bmatrix} \dot{\bar{\mathbf{x}}}_c \\ \dot{\bar{\mathbf{x}}}_{nc} \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{A}}_c & \bar{\mathbf{A}}_{12} \\ \mathbf{0} & \bar{\mathbf{A}}_{nc} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{x}}_c \\ \bar{\mathbf{x}}_{nc} \end{bmatrix} + \begin{bmatrix} \bar{\mathbf{B}}_c \\ \mathbf{0} \end{bmatrix} \mathbf{u} \tag{24.187}$$

$$\mathbf{y} = \begin{bmatrix} \bar{\mathbf{C}}_c & \bar{\mathbf{C}}_{nc} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{x}}_c \\ \bar{\mathbf{x}}_{nc} \end{bmatrix} + \mathbf{D}\mathbf{u} \tag{24.188}$$

The **controllable subspace** of a state space model is composed of all states generated through every possible linear combination of the states in $\bar{\mathbf{x}}_c$. The stability of this subspace is determined by the location of the eigenvalues of $\bar{\mathbf{A}}_c$.

On the other hand, the **uncontrollable subspace** is composed of all states generated through every possible linear combination of the states in $\bar{\mathbf{x}}_{nc}$. The stability of this subspace is determined by the location of the eigenvalues of $\bar{\mathbf{A}}_{nc}$.

Hence, the input will have no effect over the uncontrollable subspace, so the best we can hope is that this uncontrollable subspace is stable, since then the state in this subspace will go to the origin. In this case the state space model is said to be **stabilizable**.

A key feature of the descriptions (24.187) and (24.188) arises from the fact that the transfer function is given by

$$\mathbf{H}(s) = \bar{\mathbf{C}}_c(s\mathbf{I} - \bar{\mathbf{A}}_c)^{-1}\bar{\mathbf{B}}_c + \mathbf{D} \qquad (24.189)$$

Equation (24.189) says that the eigenvalues of the uncontrollable subspace do not belong to the set of poles of the system transfer function. This implies that there is a cancellation of all poles corresponding to the roots of $(s\mathbf{I} - \bar{\mathbf{A}}_{nc})$.

Controllability Canonical Form

Lemma 24.2 *Consider a completely reachable state space model for a SISO system. Then, there exists a similarity transformation which converts the state space model into the following* ***controllability canonical form:***

$$\mathbf{A}' = \begin{bmatrix} 0 & 0 & \dots & 0 & -\alpha_0 \\ 1 & 0 & \dots & 0 & -\alpha_1 \\ 0 & 1 & \dots & 0 & -\alpha_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & -\alpha_{n-1} \end{bmatrix}, \qquad \mathbf{B}' = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad (24.190)$$

where $\lambda^n + \alpha_{n-1}\lambda^{n-1} + \cdots + \alpha_1\lambda + \alpha_0 = \det(\lambda\mathbf{I} - \mathbf{A})$ *is the characteristic polynomial of* $\mathbf{A}$.

Lemma 24.3 *Consider a completely controllable state space model for a SISO system. Then, there exists a similarity transformation which converts the state space model into the following* ***controller canonical form:***

$$\mathbf{A}'' = \begin{bmatrix} -\alpha_{n-1} & -\alpha_{n-2} & \dots & -\alpha_1 & -\alpha_0 \\ 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{bmatrix}, \qquad \mathbf{B}'' = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad (24.191)$$

where $\lambda^n + \alpha_{n-1}\lambda^{n-1} + \cdots + \alpha_1\lambda + \alpha_0 = \det(\lambda\mathbf{I} - \mathbf{A})$ *is the characteristic polynomial of* $\mathbf{A}$.

Observability, Reconstructibility, and Detectability

If we consider the state space model of a system, one might conjecture that if one observes the output over some time interval then this might tell us some information about the state. The associated model property is called observability (or reconstructibility).

Observability

Observability is concerned with the issue of what can be said on the state if we measure the plant output.

Example 24.15

If we look at the system defined by state space model

$$\begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}, \qquad y(t) = [1 \quad 0] \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \tag{24.192}$$

we can see that the output $y(t)$ only is determined by $x_1(t)$, and the other state variable $x_2(t)$ has no influence on the output. So the system is not completely observable.

A formal definition is as follows:

Definition 24.3 The state $\mathbf{x_o} \neq \mathbf{0}$ is said to be **unobservable** if given $\mathbf{x}(\mathbf{0}) = \mathbf{x_o}$, and $\mathbf{u}(\mathbf{t}) = \mathbf{0}$ for $t \geq 0$, then $\mathbf{y}(t) = \mathbf{0}$ for $t \geq 0$, i.e., we cannot see any effect of $\mathbf{x_o}$ on the system output.

The system is said to be **completely observable** if there exists no nonzero initial state that it is unobservable.

Reconstructibility

There is another concept, closely related to observability, called **reconstructibility**. Reconstructibility is concerned with what can be said about $\mathbf{x}(T)$, having observed the **past** values of the output, **y**, for $0 \leq t \leq T$. For linear time invariant, continuous-time systems, the distinction between observability and reconstructibility is unnecessary. However, the following example illustrates that in discrete time, the two concepts are different. Consider

$$\mathbf{x}[t+1] = 0, \qquad \mathbf{x}[0] = \mathbf{x_o} \tag{24.193}$$

$$\mathbf{y}[t] = 0 \tag{24.194}$$

This system is clearly reconstructible for all $T \geq 1$, since we know for certain that $\mathbf{x}[T] = 0$ for $T \geq 1$. However, it is completely unobservable since $\mathbf{y}[t] = 0$, $\forall k$ irrespective of $\mathbf{x_o}$.

In view of the subtle difference between observability and reconstructibility, we will use the term *observability* in the sequel to cover the stronger of the two concepts.

Observability Test

A test for observability of a system is established in the following theorem.

Theorem 24.4 *Consider the linear, continuous, time-invariant, state space model where* $\mathbf{A} \in \mathbb{R}^{n \times n}$

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t) \tag{24.195}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t) \tag{24.196}$$

i) *The set of all unobservable states is equal to the null space of the observability matrix* $\mathbf{\Gamma}_o[\mathbf{A}, \mathbf{C}]$ where

$$\mathbf{\Gamma}_o[\mathbf{A}, \mathbf{C}] \triangleq \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \vdots \\ \mathbf{CA}^{n-1} \end{bmatrix} \tag{24.197}$$

ii) *The system is completely observable if and only if* $\mathbf{\Gamma}_o[\mathbf{A}, \mathbf{C}]$ *has **full column rank** n.*

Example 24.16

Consider the following state space model:

$$\mathbf{A} = \begin{bmatrix} -3 & -2 \\ 1 & 0 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \mathbf{C} = [1 \quad -1] \tag{24.198}$$

The observability matrix is given by

$$\mathbf{\Gamma}_o[\mathbf{A}, \mathbf{C}] = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -4 & -2 \end{bmatrix} \tag{24.199}$$

Hence rank $\mathbf{\Gamma}_o[\mathbf{A}, \mathbf{C}] = 2$, which says that the system is completely observable.

Example 24.17

If we look at the model defined in (24.192), we have

$$\mathbf{A} = \begin{bmatrix} -1 & 0 \\ 1 & -1 \end{bmatrix}, \qquad \mathbf{C} = [1 \quad 0] \tag{24.200}$$

The observability matrix is

$$\mathbf{\Gamma}_o[\mathbf{A}, \mathbf{C}] = \begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix} \tag{24.201}$$

Hence rank $\mathbf{\Gamma}_o[\mathbf{A}, \mathbf{C}] = 1 < 2$ and the system is not completely observable.

The above result also applies to discrete-time models.

The observability is a system property that does not depend on the choice of state variables. It can be proved that the rank of the matrix defined in Eq. (24.197) does not change when a similarity transformation $\boldsymbol{T}$ is used (see subsection "State Similarity Transformation").

Loss of Observability

Lack of observability may arise from structural system features. However, it is also possible that lack of observability occurs when certain system parameters take some specific numerical values. This is the same phenomenon, for controllability, we analyzed in the subsection "Controllability, Reachability, and Stabilizability." We expect that those parameters will affect the complete observability of the model in a similar way. Let us look at the following example.[2]

Example 24.18

Consider the electronic circuit in Fig. 24.14. We can see this is the same as that in Fig. 24.13 where the left and right halves were swaped, so we can use similar equations to obtain a state space model. The state variables have been chosen to be $x_1(t) = v_{C3}(t)$ and $x_2(t) = i_{R1}(t)$.

For the left half of the circuit, we have

$$\frac{dv_{C3}(t)}{dt} = -\frac{1}{R_3C_3}v_{C3}(t) + \frac{1}{R_3C_3}v_i(t) \tag{24.202}$$

$$v_+(t) = v_{C3}(t) \tag{24.203}$$

[2] Which is the *dual* of Example 24.13.

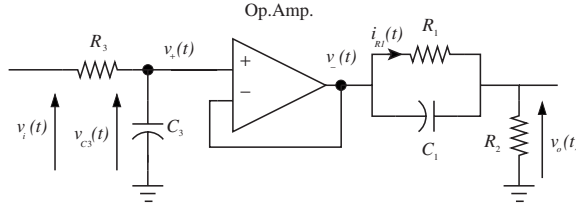



FIGURE 24.14 Electronic circuit.

And for the right half, we have

$$\frac{di_{R1}(t)}{dt} = -\left(\frac{R_1 + R_2}{C_1 R_1 R_2}\right) i_{R1}(t) + \frac{1}{C_1 R_1 R_2} v_-(t) \tag{24.204}$$

$$v_o(t) = -R_1 i_{R1}(t) + v_-(t) \tag{24.205}$$

The operational amplifier, in voltage follower connection, ensures that $v_+(t) = v_-(t)$, so we can combine the state space models given in Eqs. (24.202)–(24.205):

$$\begin{bmatrix} \frac{dv_{C3}(t)}{dt} \\ \frac{di_{R1}(t)}{dt} \end{bmatrix} = \begin{bmatrix} -\frac{1}{R_3 C_3} & 0 \\ \frac{1}{C_1 R_1 R_2} & -\frac{R_1 + R_2}{C_1 R_1 R_2} \end{bmatrix} \begin{bmatrix} v_{C3}(t) \\ i_{R1}(t) \end{bmatrix} + \begin{bmatrix} \frac{1}{R_3 C_3} \\ 0 \end{bmatrix} v_i(t) \tag{24.206}$$

$$v_o(t) = [1 \ -R_1] \begin{bmatrix} v_{C3}(t) \\ i_{R1}(t) \end{bmatrix} \tag{24.207}$$

The observability matrix is given by

$$\mathbf{\Gamma}_c[\mathbf{C}, \mathbf{A}] = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \end{bmatrix} = \begin{bmatrix} 1 & -R_1 \\ -\frac{1}{R_3 C_3} - \frac{1}{R_2 C_1} & \frac{R_1 + R_2}{R_2 C_1} \end{bmatrix} \tag{24.208}$$

To determine the complete observability, or otherwise, we need to compute the matrix determinant

$$\det(\mathbf{\Gamma}_c[\mathbf{C}, \mathbf{A}]) = \frac{1}{R_3 C_3 C_1}(-R_1 C_1 + R_3 C_3) \tag{24.209}$$

from where we conclude that the model system is completely observable if and only if, $R_1 C_1 \neq R_3 C_3$, which is the same condition we obtained in Example 24.13.

Applying Laplace transform to Eqs. (24.204)–(24.203) we obtain the transfer function from $V_i(s)$ to $V_o(s)$:

$$\frac{V_o(s)}{V_i(s)} = \frac{V_+(s)}{V_i(s)} \frac{V_o(s)}{V_-(s)} = \frac{s + \frac{1}{R_1 C_1}}{s + \frac{R_1 + R_2}{R_1 R_2 C_1}} \cdot \frac{\frac{1}{R_3 C_3}}{s + \frac{1}{R_3 C_3}} \tag{24.210}$$

The condition $R_1C_1 = R_3C_3$ produces the loss of complete observability, leading to a **pole-zero cancellation** in the model transfer function, i.e., the pole from the left half of the circuit in Fig. 24.14 is cancelled by the zero from the right half. There is subtle difference between the transfer functions in (24.210) and (24.169). The final result is the same, but the order the cancellation is different in each case. The **zero-pole cancellation** is connected to the loss of complete observability and the **pole-zero cancellation** is connected to the loss of complete controllability. These issues will be discussed in more detail in subsection "Canonical Decomposition."

Observability Gramian

The observability test in Theorem 24.4 answers *yes* or *no* to the question about completely observability of a model. However, sometimes we are interested in the *degree* of observability for a particular model. So we can quantify the energy of the output signal $\mathbf{y}(t)$, when there is no input ($\mathbf{u}(t) = \mathbf{0}$) and the state is $\mathbf{x}(0) = \mathbf{x_0}$ at $t = 0$

$$E(\mathbf{x}_0) = \int_0^\infty \|\mathbf{y}(t)\|^2 dt = \int_0^\infty \mathbf{y}(t)^T \mathbf{y}(t) dt \tag{24.211}$$

It can be proved that the *ouput energy* is

$$E(\mathbf{x}_0) = \int_0^\infty \|\mathbf{y}(t)\|^2 dt = \mathbf{x}_0^T \mathbf{Q} \mathbf{x}_0 \tag{24.212}$$

where

$$\mathbf{Q} = \int_0^\infty e^{\mathbf{A}^T t} \mathbf{C}^T \mathbf{C} e^{\mathbf{A}t} dt \tag{24.213}$$

The matrix $\mathbf{Q}$ is called **observability gramian**, and it measures the observability of the state vector $\mathbf{x}(0)$. If this matrix is *small*, it means that we have a weak contribution of the initial state $\mathbf{x}_0$ in the energy of the output $\mathbf{y}(t)$. Indeed, we can appreciate the effect of each one of the state variables taking, for example, $\mathbf{x_0} = [0,\ldots,0,1,0,\ldots,0]^T$.

Note that the existence of the integral defined in (24.213) is guaranteed if and only if the system is stable, i.e., if and only if the eigenvalues of $\mathbf{A}$ have negative real part.

Also, the observability gramian $\mathbf{Q}$ defined in (24.213) satisfies the Lyapunov equation

$$\mathbf{A}^T\mathbf{Q} + \mathbf{Q}\mathbf{A} + \mathbf{C}^T\mathbf{C} = 0 \tag{24.214}$$

For stable discrete-time systems, the controllability gramian is defined by

$$\mathbf{Q_d} = \sum_{k=0}^{\infty} (\mathbf{A_d^T})^k \mathbf{C_d^T} \mathbf{C_d} \mathbf{A_d^k} \tag{24.215}$$

which satisfies

$$\mathbf{A_d^T}\mathbf{Q_d}\mathbf{A_d} - \mathbf{Q_d} + \mathbf{C_d^T}\mathbf{C_d} = 0 \tag{24.216}$$

Example 24.19

We will use the model of Example 24.18, described by the state space models (24.206) and (24.207), to appreciate the utility of the observability gramian (24.213), especially when the model is close to losing complete observability, i.e., when $R_1C_1 \approx R_3C_3$.

Assuming the same component values as in Example 24.14 for R_1, R_2, R_3, C_1, and C_3 we have

$$\begin{bmatrix} \dot{v}_{C3}(t) \\ \dot{i}_{R1}(t) \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 10^{-3} & -\frac{20}{9} \end{bmatrix} \begin{bmatrix} v_{C3}(t) \\ i_{R1}(t) \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} v_i(t) \tag{24.217}$$

$$v_o(t) = [1 \quad -10^3] \begin{bmatrix} v_{C3}(t) \\ i_{R1}(t) \end{bmatrix} \tag{24.218}$$

If we look at the relative magnitude of the components of **C** matrix, we can foretell *a priori* that the output $v_o(t)$ will be mainly determined by state $i_{R1}(t)$. To verify this we compute the observability gramian defined in (24.172), solving

$$\mathbf{0} = \mathbf{A}^T\mathbf{Q} + \mathbf{QA} + \mathbf{C}^T\mathbf{C} \tag{24.219}$$

$$\mathbf{0} = \begin{bmatrix} -1 & 10^{-3} \\ 0 & -\frac{20}{9} \end{bmatrix} \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} + \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 10^{-3} & -\frac{20}{9} \end{bmatrix} + \begin{bmatrix} 1 \\ -10^3 \end{bmatrix} [1 \quad 10^{-3}] \tag{24.220}$$

We have

$$\mathbf{Q} = \begin{bmatrix} 0.57 & 69.83 \\ 69.83 & 225000 \end{bmatrix} \tag{24.221}$$

From there we can compute the contribution of each state to the total energy in the output. Doing this, we verify that the state variable $i_{R1}(t)$ has an effect over the output greater than the effect of $v_{C3}(t)$, as defined in Eq. (24.212):

$$\mathbf{x_0} = [1, 0]^T \Rightarrow E(\mathbf{x}_0) = 0.57 \tag{24.222}$$

$$\mathbf{x_0} = [0, 1]^T \Rightarrow E(\mathbf{x}_0) = 225000 \tag{24.223}$$

The transfer function is

$$\frac{V_o(s)}{V_i(s)} = \frac{s + 1 + \frac{1}{9}}{s + \frac{20}{9}} \cdot \frac{1}{s+1} \tag{24.224}$$

We observe that there is a pole-zero **quasi-cancellation**.

Duality Principle

We observe a remarkable similarity between the results in Theorem 24.3 and in Theorem 24.4, and also for the definitions of the gramians (24.172) and (24.213). This is known as the **duality principle**, and it can be formalized as follows:

Theorem 24.5 (Duality) *Consider a state space model described by the 4-tuple* (**A**, **B**, **C**, **D**). *Then the system is completely controllable if and only if the dual system* ($\mathbf{A}^T$, $\mathbf{C}^T$, $\mathbf{B}^T$, $\mathbf{D}^T$) *is completely observable.*

Canonical Decomposition and Detectability

The above theorem can often be used to go from a result on controllability to one on observability and vice versa. The *dual* of Lemma 24.1 is:

Lemma 24.4 *If rank* $\{\Gamma_o[\mathbf{A}, \mathbf{C}]\} = k < n$*, there exists a similarity transformation* $\mathbf{T}$ *such that with* $\bar{\mathbf{x}} = \mathbf{T}^{-1}\mathbf{x}, \bar{\mathbf{A}} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T}, \bar{\mathbf{C}} = \mathbf{C}\mathbf{T}$*, then* $\bar{\mathbf{C}}$ *and* $\bar{\mathbf{A}}$ *take the form*

$$\bar{\mathbf{A}} = \begin{bmatrix} \bar{\mathbf{A}}_o & 0 \\ \bar{\mathbf{A}}_{21} & \bar{\mathbf{A}}_{no} \end{bmatrix}, \qquad \bar{\mathbf{C}} = [\bar{\mathbf{C}}_o \quad 0] \tag{24.225}$$

where $\bar{\mathbf{A}}_o$ *has dimension k and the pair* $(\bar{\mathbf{C}}_o, \bar{\mathbf{A}}_o)$ *is completely observable.*

This result has a relevance similar to that of the controllability property and the associated decomposition. To appreciate this, we apply the dual of Lemma 24.1 to express the (transformed) state and output equations in partitioned form as

$$\begin{bmatrix} \dot{\bar{\mathbf{x}}}_o(t) \\ \dot{\bar{\mathbf{x}}}_{no}(t) \end{bmatrix} = \begin{bmatrix} \bar{\mathbf{A}}_o & 0 \\ \bar{\mathbf{A}}_{21} & \bar{\mathbf{A}}_{no} \end{bmatrix} \begin{bmatrix} \bar{\mathbf{x}}_o(t) \\ \bar{\mathbf{x}}_{no}(t) \end{bmatrix} + \begin{bmatrix} \bar{\mathbf{B}}_o \\ \bar{\mathbf{B}}_{no} \end{bmatrix} \mathbf{u}(t) \tag{24.226}$$

$$\mathbf{y}(t) = \begin{bmatrix} \bar{\mathbf{C}}_o & 0 \end{bmatrix} \begin{bmatrix} \bar{\mathbf{x}}_o(t) \\ \bar{\mathbf{x}}_{no}(t) \end{bmatrix} + \mathbf{D}\mathbf{u}(t) \tag{24.227}$$

The above description reveals why one can be in trouble when trying to control a system using only the system output. The output has no information on the state $\bar{x}_{no}$.

The **observable subspace** of a model is the space composed of all states generated through every possible linear combination of the states in $\bar{\mathbf{x}}_o$. The stability of this subspace is determined by the location of the eigenvalues of $\bar{\mathbf{A}}_o$.

The **unobservable subspace** of a model is the space composed of all states generated through every possible linear combination of the states in $\bar{x}_{no}$. The stability of this subspace is determined by the location of the eigenvalues of $\bar{\mathbf{A}}_{no}$.

If the unobservable subspace is stable we say that the system is **detectable**.

A key feature of the descriptions (24.226) and (24.227) arises from the fact that the transfer function is given by

$$\mathbf{H}(s) = \bar{\mathbf{C}}_0(s\mathbf{I} - \bar{\mathbf{A}}_o)^{-1}\bar{\mathbf{B}}_o + \mathbf{D} \tag{24.228}$$

Equation (24.228) says that the eigenvalues of the unobservable subspace do not belong to the set of poles of the system transfer function. This implies that there is a cancellation of all poles corresponding to the roots of $(s\mathbf{I} - \bar{\mathbf{A}}_{no})$.

Observability Canonical Form

There are also duals of the canonical forms given in Lemmas 24.2 and 24.3. For example, the dual of Lemma 24.3 is:

Lemma 24.5 *Consider a completely observable SISO system. Then there exists a similarity transformation that converts the model to the* ***observer canonical form:***

$$\dot{x}(t) = \begin{bmatrix} -\alpha_{n-1} & 1 & & \\ \vdots & & \ddots & \\ \vdots & & & 1 \\ -\alpha_0 & 0 & & 0 \end{bmatrix} x(t) + \begin{bmatrix} b_{n-1} \\ \vdots \\ \vdots \\ b_0 \end{bmatrix} u(t) \tag{24.229}$$

$$y(t) = [1 \quad 0 \quad \cdots \quad 0]x(t) + \mathbf{D}u(t) \tag{24.230}$$

Canonical Decomposition

Further insight into the structure of linear dynamical systems is obtained by considering those systems which are only partially observable or controllable. These systems can be separated into completely observable and completely controllable systems.

The two results of Lemmas 24.1 and 24.4 can be combined for those systems, which are neither completely observable nor completely controllable. We can see it as follows.

Theorem 24.6 (Canonical Decomposition Theorem) *Consider a system described in state space form. Then, there always exists a similarity transformation* $\mathbf{T}$ *such that the transformed model for* $\bar{\boldsymbol{x}} = \mathbf{T}^{-1}\boldsymbol{x}$ *takes the form*

$$\bar{\mathbf{A}} = \begin{bmatrix} \bar{\mathbf{A}}_{co} & \mathbf{0} & \bar{\mathbf{A}}_{13} & \mathbf{0} \\ \bar{\mathbf{A}}_{21} & \bar{\mathbf{A}}_{22} & \bar{\mathbf{A}}_{23} & \bar{\mathbf{A}}_{24} \\ \mathbf{0} & \mathbf{0} & \bar{\mathbf{A}}_{33} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \bar{\mathbf{A}}_{34} & \bar{\mathbf{A}}_{44} \end{bmatrix}, \qquad \bar{\mathbf{B}} = \begin{bmatrix} \bar{\mathbf{B}}_1 \\ \bar{\mathbf{B}}_2 \\ 0 \\ 0 \end{bmatrix}, \qquad \bar{\mathbf{C}} = [\bar{\mathbf{C}}_1 \;\; \mathbf{0} \;\; \bar{\mathbf{C}}_2 \;\; \mathbf{0}] \tag{24.231}$$

where

i) *The subsystem* $[\bar{\mathbf{A}}_{co}, \bar{\mathbf{B}}_1, \bar{\mathbf{C}}_1]$ *is both completely controllable and completely observable and has the same transfer function as the original system (see Lemma 24.6).*

ii) *The subsystem*

$$\begin{bmatrix} \bar{\mathbf{A}}_{co} & \mathbf{0} \\ \bar{\mathbf{A}}_{21} & \bar{\mathbf{A}}_{22} \end{bmatrix}, \quad \begin{bmatrix} \bar{\mathbf{B}}_1 \\ \bar{\mathbf{B}}_2 \end{bmatrix}, \quad [\bar{\mathbf{C}}_1 \;\; \mathbf{0}] \tag{24.232}$$

is completely controllable.

iii) *The subsystem*

$$\begin{bmatrix} \bar{\mathrm{A}}_{co} & \bar{\mathrm{A}}_{13} \\ 0 & \bar{\mathrm{A}}_{33} \end{bmatrix}, \quad \begin{bmatrix} \bar{\mathbf{B}}_1 \\ 0 \end{bmatrix}, \quad [\bar{\mathbf{C}}_1 \;\; \bar{\mathbf{C}}_2] \tag{24.233}$$

is completely observable.

The canonical decomposition described in Theorem 24.6 leads to an important consequence for the transfer function of the model, which will take only the completely observable and completely controllable subspace.

Lemma 24.6 *Consider the transfer function matrix* $\mathbf{H}(s)$ *given by*

$$\boldsymbol{Y}(s) = \mathbf{H}(s)\boldsymbol{U}(s) \tag{24.234}$$

Then

$$\mathbf{H} = \mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B} + \mathbf{D} = \bar{\mathbf{C}}_1(s\mathbf{I} - \bar{\mathbf{A}}_{co})^{-1}\bar{\mathbf{B}}_1 + \mathbf{D} \tag{24.235}$$

where $\bar{\mathbf{C}}_1$, $\bar{\mathbf{A}}_{co}$, *and* $\bar{\mathbf{B}}_1$ *are as in Eq. (24.231). This state description is a minimal realization of the transfer function.*

If $\mathbf{M}$ is any square matrix and we denote by $\Lambda\{\mathbf{M}\}$ the set of eigenvalues of $\mathbf{M}$, then

$$\Lambda\{\bar{\mathbf{A}}\} = \Lambda\{\bar{\mathbf{A}}_{co}\} \cup \Lambda\{\bar{\mathbf{A}}_{22}\} \cup \Lambda\{\bar{\mathbf{A}}_{33}\} \cup \Lambda\{\bar{\mathbf{A}}_{44}\} \tag{24.236}$$

where

$\Lambda\{\bar{\mathbf{A}}\}$ = eigenvalues of the system,
$\Lambda\{\bar{\mathbf{A}}_{co}\}$ = eigenvalues of the controllable and observable subsystem,
$\Lambda\{\bar{\mathbf{A}}_{22}\}$ = eigenvalues of the controllable but unobservable subsystem,
$\Lambda\{\bar{\mathbf{A}}_{33}\}$ = eigenvalues of the uncontrollable but observable subsystem,
$\Lambda\{\bar{\mathbf{A}}_{44}\}$ = eigenvalues of the uncontrollable and unobservable subsystem.

We observe that controllability for a given system depends on the structure of the input ports, i.e., where, in the system, the manipulable inputs are applied. Thus, the states of a given subsystem may be uncontrollable for a given input, but completely controllable for another. This distinction is of fundamental importance in control system design since not all plant inputs can be manipulated (consider, for example, disturbances) and, therefore, cannot be used to steer the plant to reach certain states.

Similarly, the observability property depends on which outputs are being considered. Certain states may be unobservable from a given output, but they may be completely observable from some other output. This also has a significant impact on output feedback control systems, since some states may not appear in the plant output being measured and feeded back. However, they may appear in crucial internal variables and thus be important to the control problem.

PBH Test

An alternative test for controllability and observability is provided by the following lemma known as **PBH** test.

Lemma 24.7 *Consider a state space model* ($\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$). *Then*

(i) *The system is not completely observable if and only if there exists a nonzero vector* $\mathbf{x} \in \mathbb{C}^n$ *and a scalar* $\lambda \in \mathbb{C}$ *such that*

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}, \quad \mathbf{C}\mathbf{x} = 0 \tag{24.237}$$

(ii) *The system is not completely controllable if and only if there exists a nonzero vector* $\mathbf{x} \in \mathbb{C}^n$ *and a scalar* $\lambda \in \mathbb{C}$ *such that*

$$\mathbf{x}^T\mathbf{A} = \lambda\mathbf{x}^T, \quad \mathbf{x}^T\mathbf{B} = 0 \tag{24.238}$$

24.7 State Observers

Basic Concepts

When the state variables have to be measured for monitoring, implementing control systems, or other purposes, there are hard technical and economical issues to face. Observers are a way to estimate the state variables based upon a system model, measurements of the plant output $\mathbf{y}(t)$, and measurements of the plant input $\mathbf{u}(t)$. This problem is a generalization of that of indirectly measuring a system variable using a system model and the measurement of some other easier-to-measure variable.

Observer Dynamics

Assume that the system has a state space model given by (24.42) and (24.43) with $\mathbf{D} = \mathbf{0}$ (a strictly proper system has been assumed). Then, the general structure of a classic observer for the system state is as shown in Fig. 24.15, where the matrix $\mathbf{J}$ is the **observer gain**.

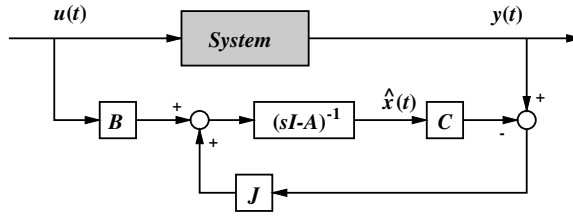



FIGURE 24.15 Classic state observer.

Therefore, the observer equation is

$$\frac{d\hat{\mathbf{x}}(t)}{dt} = \mathbf{A}\hat{\mathbf{x}}(t) + \mathbf{B}\mathbf{u}(t) + \mathbf{J}(\mathbf{y}(t) - \mathbf{C}\hat{\mathbf{x}}(t)) \tag{24.239}$$

An obvious question is: if we know an exact system model and the system input, why do we need to feed the system output? The answer is that **we need the output measurement since we do not know the system initial state**. This can be appreciated from the equation for the state estimation error, $\tilde{\mathbf{x}}(t) = \mathbf{x}(t) - \hat{\mathbf{x}}(t)$. That equation can be obtained subtracting (24.239) from (24.42). This leads to

$$\frac{d\tilde{\mathbf{x}}(t)}{dt} = (\mathbf{A} - \mathbf{J}\mathbf{C})\tilde{\mathbf{x}}(t) \tag{24.240}$$

From (24.240) we observe that the estimation error will converge to zero for a nonzero initial error if and only if all the eigenvalues of the matrix $\mathbf{A} - \mathbf{JC}$ have negative real parts, i.e., if the **observer polynomial** $E(s) = \det(sI - \mathbf{A} + \mathbf{JC})$ is strictly Hurwitz.

Discussion

- Equation (24.240) is valid only if the model is a perfect representation of the system under study. Modelling errors will impact the observer. This will normally lead to nonzero state estimation errors.
- If the pair $(\mathbf{A}, \mathbf{C})$ is completely observable, then the eigenvalues of $\mathbf{A} - \mathbf{JC}$ can be arbitrarily located (in the stability region). Thus, the speed of the estimation convergence is a designer's choice. Those eigenvalues are known as the **observer poles**.
- If the pair $(\mathbf{A}, \mathbf{C})$ is detectable, then the observer will yield zero steady state error asymptotically, although not all the eigenvalues of $\mathbf{A} - \mathbf{JC}$ can be placed at will.
- If the system is not completely observable, and the unobservable subspace contains unstable modes, then the observer will never converge.

To illustrate the observer techniques we refer to Example 24.5.

Example 24.20

Assume that we want the observer poles for the state model in Example 24.5 to be located at $s = -4$, $s = -6$, and $s = -8$. We can then compute the observer gain, **J**, using a software such as MATLAB. This yields

$$J = [-4.5247 \quad -7.5617 \quad -4.1543]^T \tag{24.241}$$

To appreciate the observer dynamics, assume that the initial system state is $\mathbf{x}(0) = [-1 \quad 2 \quad 1]^T$ and that the system input is a square wave of amplitude 1, and frequency equal to 1 rad/s. The observer is initialized with $\hat{\mathbf{x}}(0) = 0$. Then the norm of the estimation error, $\|\tilde{\mathbf{x}}(t)\|$, evolves as shown in Fig. 24.16. It is important

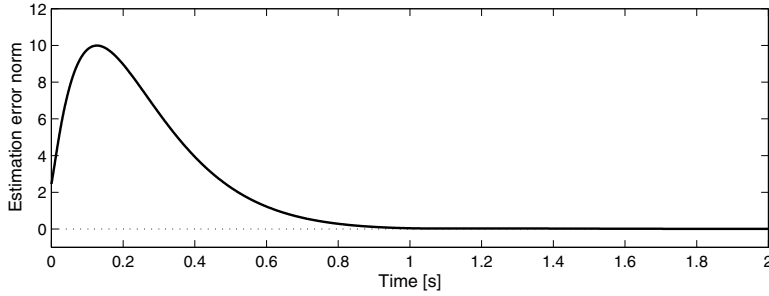



FIGURE 24.16 State estimation error.

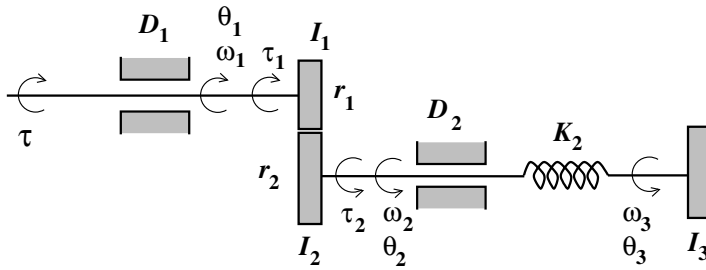



FIGURE 24.17 Rotational system.

to point out that, in this example, the plant is unstable. This means that the state and the state estimation grow unbounded. However, under the assumption of perfect modelling, the estimation error converges to zero.

To gain physical insight into the observer philosophy, we consider the following application.

Example 24.21

Figure 24.17 shows the schematics of a rotational system driven by a torque $\tau(t)$. The system power is transmitted through a gear system built with two wheels with radii r_1 and r_2 and inertias I_1 and I_2, respectively. The rotation of both shafts is damped by viscous friction with coefficients D_1 and D_2, and a significant torsional spring in shaft 2 has also been modelled. The system load is modelled as an inertia I_3. We want to estimate the load speed ω_3 based on the measurement of the speed in shaft 1, ω_1.

We first need to build a state space model. To do that we choose a minimum set of system variables, which quantify the energy stored in the system. The system has four components able to store energy: three inertias and a spring. Nevertheless, the energy stored in I_1 and I_2 can be computed either from ω_1 or from ω_2, i.e., we need only one of these speeds, since they satisfy

$$\frac{\omega_1(t)}{\omega_2(t)} = \frac{r_2}{r_1} \quad \text{and} \quad \tau_1(t)\omega_1(t) = \tau_2(t)\omega_2(t) \tag{24.242}$$

Thus, a physically oriented choice of state variables is

$$x_1(t) = \omega_1(t) \tag{24.243}$$

$$x_2(t) = \theta_2(t) - \theta_3(t) \tag{24.244}$$

$$x_3(t) = \omega_3(t) \tag{24.245}$$

From first principles we have

$$\tau(t) = D_1\omega_1(t) + I_1\frac{d\omega_1(t)}{dt} + \tau_1(t) \tag{24.246}$$

$$\tau(t) = \frac{r_2}{r_1}\tau_1(t) = D_2\omega_2(t) + I_2\frac{d\omega_2(t)}{dt} + K_2(\theta_2(t) - \theta_3(t)) \tag{24.247}$$

$$0 = K_2(\theta_3(t) - \theta_2(t)) + I_3\frac{d\omega_3(t)}{dt} \tag{24.248}$$

Since we have chosen $\omega_1(t)$ as the measurable system variable, we finally obtain

$$\frac{d\mathbf{x}(t)}{dt} = \underbrace{\begin{bmatrix} -\frac{r_1^2 D_2 + r_2^2 D_1}{r_1^2 I_2 + r_2^2 I_1} & -\frac{r_1 r_2 K_2}{r_1^2 I_2 + r_2^2 I_1} & 0 \\ \frac{r_1}{r_2} & 0 & -1 \\ 0 & \frac{K_2}{I_3} & 0 \end{bmatrix}}_{\mathbf{A}} \mathbf{x}(t) + \underbrace{\begin{bmatrix} \frac{r_2^2}{r_1^2 I_2 + r_2^2 I_1} \\ 0 \\ 0 \end{bmatrix}}_{\mathbf{B}} \tau(t) \tag{24.249}$$

$$\omega_1(t) = \underbrace{[1 \quad 0 \quad 0]}_{\mathbf{C}} \mathbf{x}(t) \tag{24.250}$$

To evaluate the observability properties of this system, numerical values for the parameters are chosen as follows:

$$r_1 = 0.25 \text{ m}, \qquad r_2 = r_3 = 0.50 \text{ m}, \qquad D_1 = D_2 = 10 \text{ Nms/rad} \tag{24.251}$$

$$K_2 = 30 \text{ Nm/rad}, \qquad I_1 = 2.39 \text{ Nms}^2\text{/rad}, \qquad I_2 = I_3 = 38.29 \text{ Nms}^2\text{/rad} \tag{24.252}$$

With these values we have that

$$\mathbf{A} = \begin{bmatrix} -1.045 & -1.254 & 0 \\ 0.5 & 0 & -1 \\ 0 & 0.784 & 0 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0.084 \\ 0 \\ 0 \end{bmatrix} \tag{24.253}$$

We next use the test presented in the subsection "Observability, Reconstructibility, and Detectability." This yields

$$\Gamma_o = \begin{bmatrix} \mathbf{C} \\ \mathbf{CA} \\ \mathbf{CA}^2 \end{bmatrix} = \begin{bmatrix} 1.0000 & 0 & 0 \\ -1.0450 & -1.2540 & 0 \\ 0.4650 & 1.3104 & 1.2540 \end{bmatrix} \tag{24.254}$$

From this expression we see that Γ_o is a full rank matrix. Thus, the system state is completely observable from $\omega_1(t)$.

Once we have a state estimate, $\hat{\mathbf{x}}(t)$, an estimate, $\hat{\omega}_3(t)$ for ω_3, is obtained from

$$\hat{\omega}_3(t) = \underbrace{[0 \quad 0 \quad 1]}_{\mathbf{K}_3^T} \hat{x}(t) \tag{24.255}$$

where $\hat{\omega}_3(t)$ can be obtained from (24.239). This yields

$$\frac{d\hat{\omega}_3(t)}{dt} = \mathbf{K}_3^T \frac{d\hat{\mathbf{x}}(t)}{dt} = \mathbf{K}_3^T(\mathbf{A} - \mathbf{JC})\hat{\mathbf{x}}(t) + \underbrace{\mathbf{K}_3^T\mathbf{B}}_{0}\ \tau(t) + \mathbf{K}_3^T\mathbf{J}\omega_1(t) \tag{24.256}$$

Observers and Measurement Noise

In the theory above we have assumed that both the system input, $\mathbf{u}(t)$, and the system output, $\mathbf{y}(t)$, are available with no errors. This assumption is usually correct with regard to $\mathbf{u}(t)$, since the same equipment generating $\mathbf{u}(t)$ is normally used to estimate the state. However, that assumption is not usually valid with respect to $\mathbf{y}(t)$, since the measurement of this variable is normally corrupted with noise. To analyze the effect of this error, let us denote by $\mathbf{y_m}(t)$ the noisy measurement, i.e., $\mathbf{y_m}(t) = \mathbf{y}(t) + \mathbf{v}(t)$, where $\mathbf{v}(t)$ is the additive measurement noise. Therefore, the state estimation error satisfies

$$\frac{d\tilde{\mathbf{x}}(t)}{dt} = (\mathbf{A} - \mathbf{JC})\tilde{\mathbf{x}}(t) + \mathbf{Jv}(t) \tag{24.257}$$

We then have that

$$\tilde{X}(s) = (s\mathbf{I} - \mathbf{A} + \mathbf{JC})^{-1}\tilde{\mathbf{x}}(0) + (s\mathbf{I} - \mathbf{A} + \mathbf{JC})^{-1}\mathbf{J}V(s) \tag{24.258}$$

Hence, the error is small if the transfer function $(s\mathbf{I} - \mathbf{A} + \mathbf{JC})^{-1}\mathbf{J}$ filters out the noise. Consider the following example.

Example 24.22

A system has a state space model given by

$$\mathbf{A} = \begin{bmatrix} -2 & 1 \\ 1 & -3 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}, \qquad \mathbf{C} = [1 \ \ -1], \qquad D = 0 \tag{24.259}$$

Assume that we want to estimate a system variable $z(t) = \gamma^T\mathbf{x}(t)$, where $\gamma^T = [1 \quad 1]$. Then, a suitable observer-based estimate is $\hat{z}(t)$, which is given by

$$\hat{z}(t) = \gamma^T\hat{\mathbf{x}}(t) \tag{24.260}$$

Then, the noise term in the estimation of $z(t)$ is $z_v(t)$, whose Laplace transform satisfies

$$Z_v(s) = H_v(s)V(s), \quad \text{where} \quad H_v(s) = \gamma^T(s\mathbf{I} - \mathbf{A} + \mathbf{JC})^{-1}\mathbf{J} \tag{24.261}$$

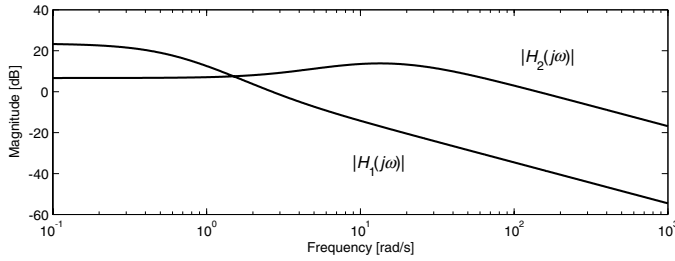



FIGURE 24.18 Observer filtering characteristics.

We next consider two different choices for the observer polynomial $E(s)$. They are

$$E_1(s) = (s+0.5)(s+0.75) \quad \text{and} \quad E_2(s) = (s+10)(s+20) \tag{24.262}$$

The reader can appreciate that the resulting observers will have very different speeds, the first observer being much slower than the second one.

With those choices we compute the observer gains, J_1 and J_2, and the corresponding filter functions

$$H_1(s) = \gamma^T(s\mathbf{I} - \mathbf{A} + \mathbf{J}_1\mathbf{C})^{-1}\mathbf{J}_1 = \frac{1.875s + 5.625}{s^2 + 1.25s + 0.375} \tag{24.263}$$

$$H_2(s) = \gamma^T(s\mathbf{I} - \mathbf{A} + \mathbf{J}_2\mathbf{C})^{-1}\mathbf{J}_2 = \frac{144s + 432}{s^2 + 30s + 200} \tag{24.264}$$

To compare both cases we compute and plot the frequency response of each filter. The result is shown in Fig. 24.18.

From Fig. 24.18 we observe that for a high frequency noise, the slowest filter is more immune to noise than the fast filter.

The above case exemplifies the trade-off between observer speed and noise immunity. A systematic way to face this dilemma is to use an optimal filter theory, such as Kalman–Bucy filtering. The interested reader is referred to [2].

24.8 State Feedback

Basic Concepts

When all the system states can be measured, and the system is completely reachable (in the sense explained in subsection "Controllability, Reactability, and Stabilizability"), we can control the system using state feedback to achieve full command of the loop dynamics. This idea is captured in Fig. 24.19.

Figure 24.19 shows the most basic form of state feedback: the plant input has a component that is proportional to the state (the other component is an external signal $\bar{r}(t)$).

State feedback is a very simple, almost naive idea. A careful analysis shows that this idea has some shortcomings and potentially dangerous features, such as

- It requires as many sensors as state variables. This is not only very expensive but also, in some cases, its implementation may become impossible.
- Each state measurement is a source of error because of its limited accuracy.
- Each measurement introduces noise, which has deleterious effect on the control system performance.

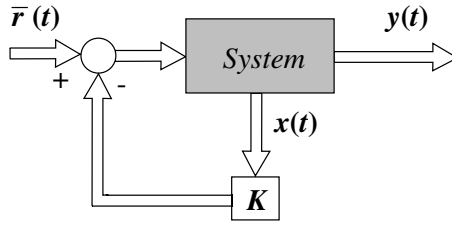



FIGURE 24.19 State feedback.

- The correct overall performance relies on the correct functioning of a complex set of equipments. This poses several questions regarding performance degradation and system integrity.

In spite of these weak points, state feedback is by itself a powerful concept, since it works as a basis for more sophisticated and robust control schemes. The key reason for this is that any linear controller can be explained as the combination of a state observer and state feedback.

Feedback Dynamics

Assume that the system to be controlled has a transfer function $\mathbf{H}(s)$ and a state space representation given by (24.42) and (24.43), with $\mathbf{D} = \mathbf{0}$. If the plant input is generated according to

$$\mathbf{u}(t) = -\mathbf{K}\mathbf{x}(t) + \bar{\mathbf{r}}(t) \tag{24.265}$$

then the state space representation for the complete control loop is given by

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{A}\mathbf{x}(t) + \mathbf{B}(-\mathbf{K}\mathbf{x}(t) + \bar{\mathbf{r}}(t)) \tag{24.266}$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) \tag{24.267}$$

It can be shown that the relationship between $\bar{\mathbf{R}}(s)$ and $\mathbf{Y}(s)$ is given by

$$\mathbf{Y}(s) = \underbrace{\mathbf{C}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B}}_{\mathbf{H}(s)}(\mathbf{I} + \mathbf{K}(s\mathbf{I} - \mathbf{A})^{-1}\mathbf{B})^{-1}\bar{\mathbf{R}}(s) \tag{24.268}$$

This shows that the state feedback loop preserves the system zeros and shifts the poles to the roots of det $(s\mathbf{I} - \mathbf{A} + \mathbf{BK})$.

Optimal State Feedback. The Optimal Regulator

Consider a linear time invariant system having a state space representation given by (24.42) and (24.43), with $\mathbf{D} = \mathbf{0}$, subject to the initial state $\mathbf{x}(0) = \mathbf{x_o}$.

Assume that the control objective is to steer the plant from the the initial state, $\mathbf{x_o}$, to the smallest possible value as soon as possible in the interval $[0, t_f]$. We additionally require that the steering process does not demand too much control effort. Then, the optimal regulator problem is defined as the problem of finding an optimal control $\mathbf{u}(t)$ over the interval $[0, t_f]$ such that a quadratic cost function is minimized. This cost function is chosen as

$$J_u(\mathbf{x_o}) = \int_0^{t_f} [\mathbf{x}(t)^T\mathbf{Q}x(t) + \mathbf{u}(t)^T\mathbf{R}\mathbf{u}(t)]dt + \mathbf{x}(t_f)^T\mathbf{Q_f}\mathbf{x}(t_f) \tag{24.269}$$

where $\mathbf{Q} \in \mathbb{R}^{n\times n}$, $\mathbf{Q}_f \in \mathbb{R}^{n\times n}$ are symmetric nonnegative definite matrices and $\mathbf{R} \in \mathbb{R}^{m\times m}$ is a symmetric positive definite matrix. The requirements on the weighting matrices are set so that the cost function makes sense. For instance, if $\mathbf{Q}$ is allowed to be negative, then the *optimal* cost could even be negative while the state could grow unbounded in magnitude. Also, if we allow $\mathbf{R}$ to have eigenvalues at the origin (i.e., $\mathbf{R}$ is allowed to be a nonnegative definite matrix, instead of requiring it to be a strictly positive definite matrix) then the control $\mathbf{u}(t)$ could also grow unbounded (in the directions of the associated eigenvectors) without that situation being revealed by the cost function.

A time invariant linear control law is asymptotically obtained when $t_f \to \infty$. Under this condition, the optimal control law is given by

$$\mathbf{u}^{\mathbf{o}}(t) = -\mathbf{K}^{\mathbf{o}}x(t) \tag{24.270}$$

with

$$\mathbf{K}^{\mathbf{o}} = -\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_\infty \tag{24.271}$$

and where $\mathbf{P}_\infty$ is the only nonnegative solution of the algebraic Riccati equation

$$\mathbf{0} = \mathbf{Q} - \mathbf{P}_\infty\mathbf{B}\mathbf{R}^{-1}\mathbf{B}^T\mathbf{P}_\infty + \mathbf{P}_\infty\mathbf{A} + \mathbf{A}^T\mathbf{P}_\infty \tag{24.272}$$

For this solution to exist, it is necessary that certain technical conditions are satisfied (for a detailed discussion of these issues see, for instance, [5]).

Discussion

- The solution for the LQR problem minimizes the cost function (24.269) and, when $t_f \to \infty$, always stabilizes the plant.
- A key issue is how to choose the weighting matrices $\mathbf{Q}$ and $\mathbf{R}$. A frequent choice for $\mathbf{Q}$ is $\mathbf{Q} = \mathbf{C}^T\mathbf{C}$. With this choice, the magnitude of the plant output is directly introduced into the cost function.
- For a given $\mathbf{Q}$, the *size* of $\mathbf{R}$ strongly influences the location of the closed loop poles. The larger $\mathbf{R}$ is, the slower is the control loop.

Further reading on optimal quadratic regulators can be found in the literature. See, e.g., [1,3,4,8,9].

24.9 Observed State Feedback

Separation Strategy

Due to the drawbacks inherent in the measuring of the state, feedback of the estimated state can be used instead. The resulting control system integrates an observer and a feedback mechanism for the observed states.

The combination of a state observer and the feedback of the estimated state conform the structure shown in Fig. 24.20.

In Fig. 24.20, the (matrix) transfer functions $\mathbf{T}_1(s)$ and $\mathbf{T}_2(s)$ can be obtained from Fig. 24.15. This yields

$$\mathbf{T}_1(s) = (s\mathbf{I} - \mathbf{A}_\mathbf{o} + \mathbf{J}\mathbf{C}_\mathbf{o})^{-1}\mathbf{B}_\mathbf{o} \tag{24.173}$$

$$\mathbf{T}_2(s) = (s\mathbf{I} - \mathbf{A}_\mathbf{o} + \mathbf{J}\mathbf{C}_\mathbf{o})^{-1}\mathbf{J} \tag{24.174}$$

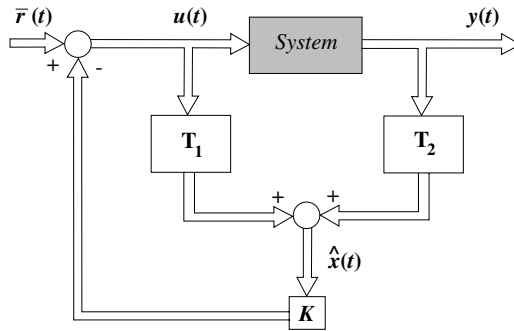



FIGURE 24.20 Estimated state feedback.

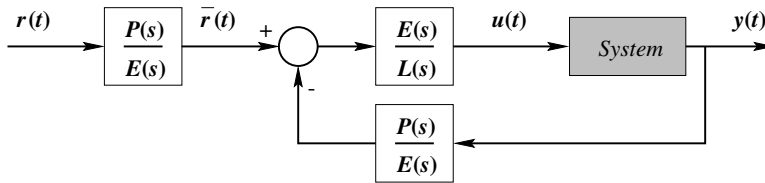



FIGURE 24.21 Equivalent control loop.

Transfer Function Interpretation for the Single-Input Single-Output Case

Consider a SISO plant having transfer function

$$G_o(s) \;=\; \mathbf{C}(s\mathbf{I} - \mathbf{A_o})^{-1}\mathbf{B} = \frac{N_o(s)}{M_o(s)} \tag{24.275}$$

where $M_o(s)$ and $N_o(s)$ are polynomials in s.

First, a state feedback gain, **K**, is chosen to obtain a closed loop polynomial $F(s)$, where $F(s) = \det(s\mathbf{I} - \mathbf{A_o} + \mathbf{B_o}\mathbf{K})$. Next, an observer gain, **J**, is computed to obtain an observer polynomial $E(s) = \det(s\mathbf{I} - \mathbf{A_o} + \mathbf{J}\mathbf{C_o})$.

If the observer and the observed state feedback are combined, the resulting control loop can be made equivalent (by a suitable choice of $\bar{r}(\mathrm{t})$) to the classical control loop shown in Fig. 24.21.

In Fig. 24.21 the polynomials $P(s)$ and $L(s)$ satisfy the Diophantine equation

$$M_o(s)L(s) + N_o(s)P(s) \;=\; E(s)F(s) \tag{24.276}$$

This result says that the set of closed loop poles is the union of the set of observer poles and the set of state feedback poles.

References

1. Anderson, B.D.O. and Moore, J., *Linear optimal Control*. Prentice-Hall, Englewood Cliffs, NJ, 1971.
2. Anderson, B.D.O. and Moore, J., *Optimal Filtering*. Prentice-Hall, Englewood Cliffs, NJ, 1979.
3. Athans, M. and Falb, P., *Optimal Control*. McGraw HIll, 1966.
4. Dennis Bernstein and Wassim Haddad. LQG control with an H_∞ performance bound: A Riccati equation approach. *IEEE Transactions on Automatic Control,* 34(3): L293–305, 1989.
5. Bittanti, S., Laub, A.J., and Willems, J.C., *The Riccati Equation*. Springer Verlag, Berlin, 1996.
6. Dorf, R.C. and Bishop, R., *Modern Control Systems*. Prentice-Hall, Englewood Cliffs, NJ, 1997.

7. Franklin, G.F. and Powel, J.D., *Digital Control of Dynamics Systems*. 2nd ed., Addison-Wesley, 1990.
8. Goodwin, G.C., Graebe, S., and Salgado, M.E., *Control System Design*. Prentice-Hall, NJ, 2001.
9. Kwakernaak, H. and Sivan, R., *Linear Optimal Control System*. Wiley–Interscience, New York, 1972.
10. Ogata, K., *State Space Analysis of Control Systems*. Prentice-Hall, Englewood Cliffis, NJ, 1967.
11. Rosenbrock, H.H., *State Space and Multivariable Theory*. John Wiley and Sons, New York, 1970.
12. Schultz, D.G. and Melsa, J.L., *State Function and Linear Control Systems*. McGraw Hill, New York, 1967.
13. Wiberg, D.W., *Theory and Problems of State Space and Linear Systems*. McGraw Hill, New York, 1971.
14. Zadeh, L.A. and Desoer, C.A., *Linear System Theory: The State Space Approach*. McGraw Hill, New York, 1963.
15. Zhou, K., *Essentials of Robust Control*. Prentice-Hall, Englewood Cliffs, NJ, 1998.
16. Zhou, K., Salomon, G., and Wu, E., Balanced realization and model reduction for unstable systems. *International Journal of Robust and Nonlinear Control,* 9:183–198, 1999.

25

Response of Dynamic Systems

Raymond de Callafon

University of California

25.1 System and Signal Analysis

In dynamic system design and analysis it is important to predict and understand the dynamic behavior of the system. Examining the dynamic behavior can be done by using a mathematical model that describes the relevant dynamic behavior of the system in which we are interested. Typically, a model is formulated to describe either continuous or discrete time behavior of a system. The corresponding equations that govern the model are used to predict and understand the dynamic behavior of the system.

A rigorous analysis can be done for relatively simple models of a dynamic system by actually computing solutions to the equations of the model. Usually, this analysis is limited to linear first and second order models. Although limited to small order models, the solutions tend to give insight in the typical responses of a dynamic system. For more complicated, higher order and possibly nonlinear models, numerical simulation tools provide an alternative for the dynamic system analysis.

In the following we review the analysis of linear models of discrete and continuous time dynamic systems. The equations that describe and relate continuous and discrete time behavior are presented. For the analysis of continuous time systems extensive use is made of the Laplace transform that converts linear differential equations into algebraic expressions. For similar purposes, a z-transform is used for discrete time systems.

Continuous Time Systems

Models that describe the linear continuous time dynamical behavior of a system are usually given in the form of differential equations that relate an input signal $u(t)$ to an output signal $y(t)$. The differential equation of a time invariant linear continuous time model has the general format

$$\sum_{j=0}^{n_a} a_j \frac{d^j}{dt^j} y(t) = \sum_{j=0}^{n_b} b_j \frac{d^k}{dt^j} u(t) \tag{25.1}$$

in which a linear combination is taken using the jth order time derivatives d^j/dt^j of a single output $y(t)$ and a single input $u(t)$. In (25.1), the scalar real valued numbers a_j for $j = 0,\dots n_a$, $a_{n_a} \neq 0$ and b_j for $j = 0,\dots,n_b$, $b_{n_b} \neq 0$, respectively, are called the denominator and numerator coefficients. The input $u(t)$ is distinguished from the output $y(t)$ in (25.1) by requiring $n_a \geq n_b$. As a result, the n_ath derivative is the highest derivative of the output $y(t)$ and n_a is used to indicate the order of the differential equation.

An alternative representation of a model of a continuous time system can be obtained by rewriting the n_ath order differential equation in (25.1) into a set of (coupled) first order differential equations. This can be done by introducing a state variable $x(t)$ and rewriting the higher order differential equation into

$$\begin{aligned} \frac{d}{dt}x(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{aligned} \tag{25.2}$$

where A, B, C, and D are real valued matrices. The set of first order differential equations given in (25.2) is referred to as a state space representation. The state variable $x(t)$ is a column vector and contains n_a variables, where n_a is the order of the differential equation.

The size of the matrices in (25.2) corresponds to the order of differential equation from which the state space realization is derived. For generalization purposes, consider multiple inputs and outputs rearranged in $m \times 1$ input column vector $u(t)$ and a $p \times 1$ output column vector $y(t)$. Given the $n_a \times 1$ size of the state vector, the state matrix A has size $n_a \times n_a$, the input matrix has size $n_a \times m$, the output matrix C has size $p \times n_a$, and the feedthrough matrix D has size $m \times p$. From these size considerations it can be observed that the state space realization in (25.2) easily generalizes the model description of multi-input multi-output systems.

To illustrate the concepts, consider the differential equation

$$m\frac{d^2}{dt^2}y(t) + c\frac{d}{dt}y(t) + ky(t) = u(t) \tag{25.3}$$

that describes the dynamical behavior of the one cart system given in Fig. 25.1. The differential equation (25.3) is found by writing Newton's second law for the cart mass m with position output $y(t)$, spring force $ky(t)$, damper force $c(d/dt)y(t)$, and force input $u(t)$. Comparing with (25.1) it can be seen that $n_a = 2 \geq n_b = 0$, making (25.3) a second order differential equation. The differential equation can be rewritten into a state space representation (25.2) by defining the state variable

$$x(t) := \begin{bmatrix} y(t) \\ \frac{d}{dt}y(t) \end{bmatrix}$$

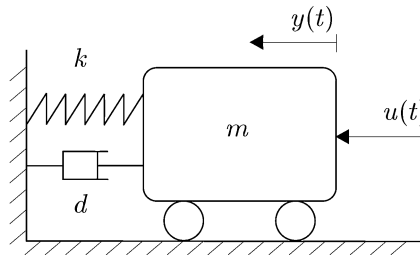



FIGURE 25.1 One cart system representing a single mass dynamical system with cart mass m, spring constant k, and damping constant c.

that consists the position and velocity of the mass. With this state variable (25.3) can be rewritten into

$$\frac{d}{dt}x(t) = \begin{bmatrix} 0 & 1 \\ -\frac{k}{m} & -\frac{d}{m} \end{bmatrix} x(t) + \begin{bmatrix} 0 \\ \frac{1}{m} \end{bmatrix} u(t)$$

$$y(t) = \begin{bmatrix} 1 & 0 \end{bmatrix} x(t) + 0u(t)$$

which yields a state space model similar to (25.2). In this case, the size of the state matrix A is 2×2, the input matrix B is 2×1, the output matrix C is 1×2, and the feedthrough matrix $D = 0$ is scalar.

Discrete Time Systems

Discrete time models approximate and describe the sampled data behavior of a continuous time dynamical system. In some applications, such as digital control, the dynamical control system is inherently discrete time. In these situations, analysis with discrete time equivalent models in necessary.

For analysis purposes, both input $u(t)$ and output $y(t)$ are assumed to be sampled on a regular discrete time interval

$$t = k\Delta T, \qquad k = 0,1,2,\ldots$$

where ΔT indicates the sampling time. To maintain uniform notation throughout the analysis, the sampling time ΔT is normalized to $\Delta T = 1$ and the time dependency t is assumed to be discrete with $t = k = 0, 1, 2, \ldots$.

Given sampled or discrete time input/output data, a linear discrete time model can be formulated in the form of a difference equation

$$\sum_{j=0}^{n_c} c_k y(k+j) = \sum_{j=0}^{n_d} d_k u(k+j) \tag{25.4}$$

in which a linear combination is taken of positive time shifted inputs $u(k)$ and outputs $y(k)$. To distinguish the differential equation from the differential equation (25.1), different scalar real valued numbers c_j for $j = 0,\ldots,n_c$, $c_{n_c} \neq 0$ and d_j for $j = 0,\ldots,n_d$, $d_{n_d} \neq 0$ are used. The input $u(k)$ is distinguished from the output $y(k)$ in (25.1) by requiring $n_c \geq n_d$ for causality purposes. As a result, the n_c is the largest time shift of the output $y(k)$ and n_c is used to indicate the order of the difference equation.

The simplicity with which the difference equation can be represented also allows an algebraic representation of (25.4). Introducing the time shift operator

$$qu(k) := u(k+1) \tag{25.5}$$

allows (25.4) to be rewritten into the algebraic expression

$$y(k)\sum_{j=0}^{n_c} c_j q^j = u(k)\sum_{j=0}^{n_d} d_j q^j$$

Following this analysis, the discrete time output $y(k)$ can be represented by the difference model

$$y(k) = G(q)u(k), \qquad \text{with } G(q) = \frac{\sum_{j=0}^{n_d} d_j q^j}{\sum_{j=0}^{n_c} c_j q^j} \tag{25.6}$$

where the scalar real valued numbers c_j for $j = 0,\ldots,n_c$, $c_{n_c} \neq 0$ and d_j for $j = 0,\ldots,n_d$, $d_{n_d} \neq 0$, respectively, indicate the denominator and numerator coefficients.

Similar to the continuous time system representation, the higher order difference equation (25.4) can also be rewritten into a set of (coupled) first order difference equations for analysis purposes. This can be done by introducing a state variable $x(k)$ and rewriting the higher order difference equation into

$$\begin{aligned} qx(k) &= Fx(k) + Gu(k) \\ y(k) &= Hx(k) + Ju(k) \end{aligned} \tag{25.7}$$

where $qx(k) = x(k + 1)$, according to (25.5). The state variable $x(k)$ is a column vector and contains n_c variables, where n_c is the order of the difference equation. The state space matrices in (25.7) are labeled differently to distinguish them from the continuous time state space model.

Laplace and *z*-Transform

An important mathematical concept for the analysis of models described by linear differential equations such as (25.1) and (25.2) is the Laplace transform. As indicated before, the Laplace transform converts linear differential equations into algebraic expressions. With this conversion, proper algebraic manipulation can be used to recover solutions of the differential equation. In a similar manner, the z-transform is used for discrete time models described by difference equations. Although it was shown in (25.6) that a difference equation can be written as an algebraic expression, the z-transform allows complex analysis of the discrete time models.

The Laplace transform of a signal $u(t)$ is defined to be

$$L\{u(t)\} := u(s) = \int_{t=0}^{\infty} u(t)e^{-st}\,dt \tag{25.8}$$

where the integration over t eliminates the time dependency and the transform $u(s)$ is a function of the Laplace variable only. This is indicated in the transform $u(s)$ where the dependency of t has been dropped, and $u(s)$ is a function of the (complex valued) Laplace variable s only.

The integral (25.8) exists for most commonly used signals $u(t)$, provided certain conditions on s are imposed. To illustrate the transform, consider a (unity) step signal

$$u(t) := \begin{cases} 0, & t < 0 \\ 1, & t \geq 0 \end{cases}$$

where the shape of $u(t)$ resembles a stepwise change of an input signal. With the definition of the Laplace transform in (25.8) the transform of the step signal becomes

$$u(s) = \int_{t=0}^{\infty} u(t)e^{-st}\,dt = \int_{t=0}^{\infty} e^{-st}\,dt = -\frac{e^{-st}}{s}\bigg|_0^{\infty} = \frac{1}{s} \tag{25.9}$$

where it is assumed that the real part of s is greater than zero so that $\lim_{t\to\infty} e^{-st} = 0$.

If a signal $u(k)$ is given at discrete time samples $k = 0, 1, 2\ldots$, the integral expression of (25.8) cannot be applied. Instead, a transform similar to the Laplace transform can be used and denoted by the z-transform. The z-transform of a discrete time signal $u(k)$ is defined as

$$L\{u(k)\} := u(z) = \sum_{k=0}^{\infty} u(k)z^{-k} \tag{25.10}$$

The series (25.10) converges if it is assumed that there exist values r_l and r_u with $r_l < |z| < r_u$ as bounds on the magnitude of the complex variable z.

The z-transform has the same role in discrete time systems that the Laplace transform has in continuous time systems. In case of sampling, the complex variable z of the z-transform is related to the complex variables s in the Laplace transform via

$$z = e^{s\Delta T} \tag{25.11}$$

where ΔT is the sampling time used for sampling. Both the Laplace and z-transform are linear operators and satisfy

$$L\{\alpha u(t) + \beta y(t)\} = \alpha L\{u(t)\} + \beta\{y(t)\} \tag{25.12}$$

Using the definition in (25.8) and the linearity property in (25.12), the transform of most commonly used functions has been precalculated and tabulated.

Of particular interest for the analysis of linear differential equations such as (25.1) and (25.2) is the Laplace transform of a derivative:

$$\begin{aligned} L\left\{\frac{d}{dt}u(t)\right\} &= \int_{t=0}^{\infty} \frac{d}{dt}u(t)e^{-st}\,dt \\ &= u(t)e^{-st}\Big|_0^{\infty} + s\int_{t=0}^{\infty} u(t)e^{-st}\,dt \\ &= su(s) - u(0) \end{aligned}$$

With $u(0) = 0$ it can be seen that the Laplace transform of the derivative of $u(t)$ is simply s times the Laplace transform of $u(s)$. This result can be extended to higher order derivatives and the result for the nth derivative is given by

$$L\left\{\frac{d^n}{dt^n}u(t)\right\} = s^n u(s) - \sum_{j=1}^{n} s^{n-j} \frac{d^{j-1}}{dt^{j-1}}u(t)\Bigg|_{t=0}$$

In case the signal $u(t)$ satisfies the initial zero conditions

$$\frac{d^{j-1}}{dt^{j-1}}u(t)\Bigg|_{t=0} = 0 \quad \text{for } j = 1, \ldots, n$$

the formula reduces to

$$L\left\{\frac{d^n}{dt^n}u(t)\right\} = s^n u(s)$$

and the Laplace transform of an nth order derivative is simply s^n times the transform $u(s)$.

For discrete time systems the interest lies in the z-transform of a time-shifted signal. Similar to the Laplace transform, the z-transform of an n time-shifted signal can be computed and is given by

$$L\{q^n u(k)\} = z^n u(z) - \sum_{j=0}^{n-1} z^{n-j} u(j)$$

In case the discrete time signal $u(k)$ satisfies the initial zero conditions $u(j) = 0$ for $j = 0,\ldots,n-1$, the formula reduces to

$$L\{q^n u(k)\} = z^n u(z)$$

and the z-transform of an n time-shifted discrete time signal is simply z^n times the transform $u(z)$.

Transfer Function Models

The results of the Laplace and z-transform can be used to reduce linear differential equations (25.1) and difference equation (25.4) to the algebraic expressions. Starting with the differential equations for continuous time models and assuming zero initial conditions for both the input $u(t)$ and output signal $y(t)$, the Laplace transform of (25.1) yields

$$y(s)\sum_{j=0}^{n_a} a_j s^j = u(s)\sum_{j=0}^{n_b} b_j s^j$$

which can be written in transfer function format

$$y(s) = G(s)u(s), \quad \text{with } G(s) = \frac{\sum_{j=0}^{n_b} b_j s^j}{\sum_{j=0}^{n_a} a_j s^j} \tag{25.13}$$

In (25.13), the transfer function $G(s)$ is the ratio of the numerator polynomial $\sum_{j=0}^{n_b} b_j s^j$ and the denominator polynomial $\sum_{j=0}^{n_a} a_j s^j$. As indicated before, the scalar real valued numbers a_j for $j = 0,\ldots,n_a$, $a_{n_a} \neq 0$ and b_j for $j = 0, \ldots, n_b$, $b_{n_b} \neq 0$, respectively, are called the denominator and numerator coefficients.

Similarly for the discrete time model, assuming zero initial conditions for both the input $u(k)$ and output signal $y(k)$, the z-transform of (25.4) yields

$$y(z)\sum_{j=0}^{n_c} c_j z^j = u(z)\sum_{j=0}^{n_d} b_j z^j$$

which can be written in transfer function format

$$y(z) = G(z)u(z), \quad \text{with } G(z) = \frac{\sum_{j=0}^{n_c} c_j z^j}{\sum_{j=0}^{n_a} a_j z^j} \tag{25.14}$$

From the transfer function representations, poles and zeros of the dynamic system can be computed for dynamic system analysis. The poles of the system are defined as the roots of the denominator polynomial. The zeros of the system are defined as the roots of the numerator polynomial.

The Laplace and z-transform can also be used to reduce the state space representation to a set of algebraic expressions that consists of (coupled) first order polynomials. Assuming zero initial conditions for the state vector $x(t)$, application of the Laplace transform to (25.2) yields

$$\begin{aligned} sx(s) &= Ax(s) + Bu(s) \\ y(s) &= Cx(s) + Du(s) \end{aligned}$$

in which the state vector $x(s)$ can be eliminated. Solving for $x(s)$ gives $x(s) = (sI - A)^{-1} Bu(s)$ and the above transform can be rewritten into a transfer function representation

$$y(s) = G(s)u(s), \quad \text{with } G(s) = D + C(sI - A)^{-1}B \tag{25.15}$$

Under mild technical conditions involving controllability and observability of the state space model, the transfer function representations in (25.13) and (25.15) are similar in case the state space model in (25.2) is derived from the differential equation (25.1) and vice versa.

25.2 Dynamic Response

The Laplace and z-transform offer the possibility to compute the dynamic response of a dynamic system by means of algebraic manipulations. The analysis of the dynamic response gives insight into the dynamic behavior of the system by addressing the response to typical test signals such as impulse, step, and sinusoid excitation of the system.

The response can be computed for relatively simple continuous or discrete dynamical systems given by low order differential or difference equations. Both the state space model and the transfer function descriptions provide helpful representations in the analysis of a dynamic system. The result are presented in the following.

Pulse and Step Response

A possible way to evaluate the response of a dynamic system is by means of pulse and step based test signals. For continuous time systems an input impulse signal is defined as a δ function

$$u_{\mathrm{imp}}(t) := \delta(t) = \begin{cases} \infty, & t = 0 \\ 0, & t \neq 0 \end{cases}$$

with the property

$$\int_{t=-\infty}^{\infty} f(t)\delta(t) = f(0)$$

where $f(t)$ is an integrable function over $(-\infty, \infty)$. Although an impulse signal is not practical from an experiment point of view, the computation or simulation of the impulse response gives insight into the transient behavior of the dynamical system.

With the properties of the impulse function $\delta(t)$ mentioned above, the Laplace transform of the impulse function is given by

$$L\{\delta(t)\} = \delta(s) = \int_{t=0}^{\infty} \delta(t)e^{-st}\,dt = e^{-s0} = 1$$

Hence the output $y(s)$ due to an impulse input is given by $y_{\mathrm{imp}}(s) = G(s)u_{\mathrm{imp}}(s) = G(s)\delta(s) = G(s)$. As a result, an immediate inverse Laplace transform of the continuous time transfer function $G(s)$,

$$y_{\mathrm{imp}}(t) = L^{-1}\{G(s)\}$$

gives the dynamic response $y_{\mathrm{imp}}(t)$ of the system to an impulse input response.

The computation of the step response is done in a similar way. In (25.9), the Laplace transform of the step signal

$$u_{\mathrm{step}}(t) := \begin{cases} 0, & t < 0 \\ 1, & t \geq 0 \end{cases}$$

is given as $u_{\text{step}}(s) = 1/s$. Consequently, with $y_{\text{step}}(s) = G(s)u_{\text{step}}(s) = G(s)/s$, the inverse Laplace transform of $G(s)/s$

$$y_{\text{step}}(t) = L^{-1}\left\{\frac{G(s)}{s}\right\}$$

will yield the dynamic response $y_{\text{step}}(t)$ of the system to a step input response.

From a practical point of view, the computation of an inverse Laplace transform is limited to low order models of first or second order. However, the results give insight into the dominant behavior of most dynamic systems. This is illustrated in the following examples.

- Consider a first order continuous model given by the transfer function

$$G(s) = \frac{K}{\tau s + 1}$$

  where K and τ indicate, respectively, the static gain and the time constant of the system. Such a transfer function may arise from a simple RC network with $\tau = \text{RC}$. In order to compute the step response of the system, the inverse Laplace transform of $G(s)/s$ needs to be computed. This inverse Laplace transform is given by

$$y_{\text{step}}(t) = L^{-1}\left\{\frac{G(s)}{s}\right\} = \frac{K}{\tau}(1 - e^{-t/\tau})$$

  and it can be seen that the step response is an exponential function. For stability the time constant τ needs to satisfy $\tau > 0$. It can also be observed that the smaller the time constant, the faster the response.

- Consider a second order continuous time model given by the transfer function

$$G(s) = \frac{\omega_n^2}{s^2 + 2\beta\omega_n s + \omega_n^2} \tag{25.16}$$

  where ω_n and β, respectively, indicate the undamped resonance frequency and the damping coefficient of the system. This model can be derived from the dynamical behavior of the one cart system depicted in Fig. 25.1 and given in (25.3). For $\beta < 1$ (underdamped), the inverse Laplace transform of $G(s)$ is given by

$$y_{\text{imp}}(t) = \frac{\omega_m}{\sqrt{1-\beta^2}} e^{-\beta\omega_n t} \sin\left(\omega_n\sqrt{1-\beta^2}\, t\right)$$

  From this expression it can be observed that the response is a decaying sinusoid with a resonance frequency of $\omega_n\sqrt{1-\beta^2}$. For stability, both $\omega_n > 0$ and $\beta > 0$ and the larger ω_n, the faster the decay of the sinusoid and the higher is the frequency of the response $y_{\text{imp}}(t)$. Illustration of the impulse response of this second order system have been depicted in Figs. 25.2 and 25.3 where variations in the undamped resonance frequency ω_n and the damping coefficient β illustrate the dynamic behavior of the system.

For discrete systems, the analysis of the pulse response is based on the discrete time pulse function

$$u_{\text{imp}}(k) := \delta(k) = \begin{cases} 1, & k = 0 \\ 0, & k \neq 0 \end{cases}$$

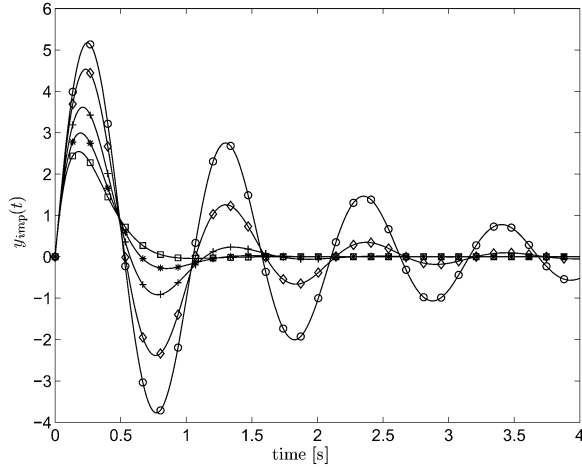



FIGURE 25.2 Variations in impulse response $y_{imp}(t)$ of second order system with $\omega_n = 6$ and $\beta = 0.1(\bigcirc)$, $0.2(\diamond)$, $0.4(+)$, $0.6\ (*)$, $0.8(\square)$.

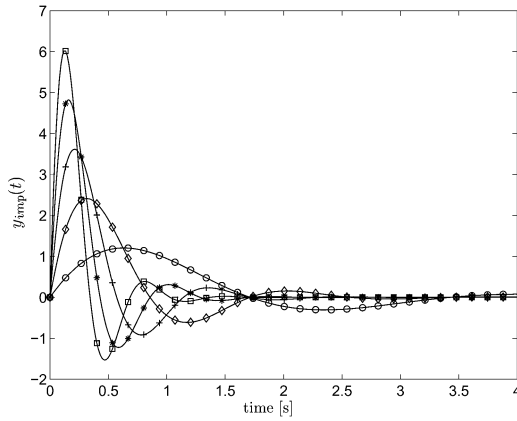



FIGURE 25.3 Variations in impulse response $y_{imp}(t)$ of second order system with $\beta = 0.4$ and $\omega_n = 2(\bigcirc)$, $4(\diamond)$, $6(+)$, $8(*)$, $10(\square)$.

which has a value of 1 at $k = 0$ and zero anywhere else. The step signal is similar to the continuous time signal and is given by

$$u_{\text{step}}(k) := \begin{cases} 0, & k < 0 \\ 1, & k \geq 1 \end{cases}$$

In order to characterize the discrete time pulse and step response a similar procedure as for the continuous time model can be followed by using the z-transform. It is easy to show that the z-transform $u_{\text{imp}}(z) = 1$ and the z-transform of the step signal equals $u_{\text{step}}z = z/(z - 1)$. Hence, the response of the discrete time system to a pulse or step signal can be computed with

$$y_{\text{imp}}(k) = L^{-1}\{G(z)\}, \qquad y_{\text{step}}(k) = L^{-1}\left\{\frac{G(z)z}{z-1}\right\}$$

In addition to the approach using a z-transform, the ratio of the polynomials in the difference model (25.6) can be written in a series expansion:

$$G(q) = \frac{\sum_{j=0}^{n_d} d_j q^j}{\sum_{j=0}^{n_c} c_j q^j} = \sum_{j=0}^{\infty} g_k q^{-k}$$

With the discrete time pulse function $u_{\text{imp}}(k)$ as an input, it can be observed that

$$y_{\text{imp}}(k) = \sum_{j=0}^{\infty} g_k q^{-k} \delta(k) = g_k$$

and it can be concluded that the pulse response $y_{\text{imp}}(k)$ equals the coefficients in the series expansion of the difference equation. Similarly, with the discrete time step function $u_{\text{step}}(k)$ as an input, it can be observed that

$$y_{\text{imp}}(k) = \sum_{j=0}^{\infty} g_k q^{-k} u_{\text{step}}(k) = \sum_{j=0}^{k} g_k$$

and it can be concluded that the step response $y_{\text{step}}(k)$ values are computed as a finite sum of the coefficients in the series expansion of the difference equation. The computation of a discrete time pulse response for a first order discrete time model is given in the following example.

- Consider a first order discrete model given by the difference model

  $$G(q) = \frac{1}{q+d}$$

  where d indicates the discrete time constant of the system. The series expansion of the difference model can be computed as follows:

  $$G(q) = \frac{1}{q-d} = \sum_{j=0}^{\infty} d^j$$

  and it can be seen that the discrete time pulse response

  $$y_{\text{imp}}(k) = d^k$$

  is an exponential function. For stability the discrete constant d needs to satisfy $|d| < 1$. Similar as in the continuous time model it can be observed that the smaller the time constant, the faster the response. Additionally, the first order discrete time model may exhibit an oscillation in case $-1 < d < 0$.

Sinusoid and Frequency Response

So far we have considered transient effects caused by step, pulse, and impulse inputs to investigate the dynamic properties of a dynamical system. However, periodic inputs occur frequently in practical situations and the analysis of a dynamic system to periodic inputs and especially sinusoidal inputs can yield more insight into the behavior of the system.

The response of a linear system to a sinusoidal input is referred to as the frequency response of the system. An input signal, $u(t) = U\sin\omega t$, that is, a sine wave with amplitude U and frequency ω_j, has a Laplace transform

$$u(s) = \frac{U\omega}{s^2+\omega^2}.$$

Consequently, the response of the system is given by

$$y(s) = G(s)\frac{U\omega}{s^2+\omega^2}$$

and a partial fraction expansion of $y(s)$ will result in terms that represent the (stable) transient behavior of $y(s)$ and the term associated to the sinusoidal input $u(s)$. Elimination of the transient effects and performing an inverse Laplace transform will yield a periodic time response $y(t)$ of the same frequency ω_j given by

$$y(t) = AU\sin(\omega t + \phi)$$

where the amplitude magnification A and the phase shift ϕ are given by

$$A = G(s)|_{s=i\omega}, \qquad \phi = \angle G(s)|_{s=i\omega} \tag{25.17}$$

By evaluating the transfer function $G(s)$ along the imaginary axis $s = i\omega$, $\omega \geq 0$, the magnitude $|G(i\omega)|$ gives information on the relative amplification of the sinusoidal input, whereas the phase $\angle G(i\omega)$ gives information on the relative phase shift between input and output.

This analysis can be easily extended to discrete time systems by employing the relation between the Laplace variable s and the z-transform variable in (25.11) to obtain the discrete time sinusoidal response

$$y(k) = AU\sin(\omega k + \phi)$$

where the amplitude magnification A and the phase shift ϕ are given by

$$A = G(z)|_{z=e^{i\Delta T\omega}}, \qquad \phi = \angle G(z)|_{z=e^{i\Delta T\omega}} \tag{25.18}$$

Due to the sampling nature of the discrete time system, the transfer function $G(z)$ is now evaluated on the unit circle

$$e^{i\Delta T\omega}, \quad 0 \leq \omega < \frac{\pi}{\Delta T}$$

to attain information of the magnitude and phase shift of the sinusoidal response.

Plotting the frequency response of a dynamical system gives insight in the pole locations (resonance modes) and zero locations of the dynamical system. As an example, the frequency response of the second order system given in (25.16) has been depicted in Fig. 25.4. It can be seen from the figure that, as expected, the second order system is less damped for smaller damping coefficients β and this results in a larger amplitude response of the second order system at the resonance frequency $\omega_n = 6$ rad/s. It can also be observed that the phase change at the resonance frequency becomes more abrupt for smaller damping coefficients.

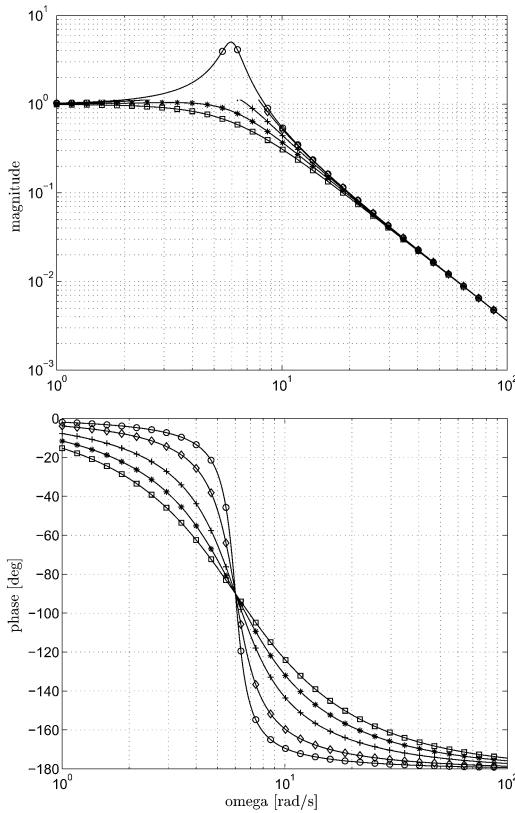



FIGURE 25.4 Variations in frequency response of second order system $G(s)$ with $\omega_n = 6$ and $\beta = 0.1(\bigcirc)$, $0.2(\Diamond)$, $0.4(+)$, $0.6(*)$, $0.8(\square)$.

25.3 Performance Indicators for Dynamic Systems

Step Response Parameters

Specifications for dynamic systems often involve requirements on the transient behavior of the system. Transient behavior requirements can be formulated on the basis of a step response and the most significant parameters have been summarized below and illustrated in Fig. 25.5.

- Steady state or DC value y_s of step response output.
- The steady state error y_{se} is the error between steady state value y_s and desired DC value of step response output.
- The maximum overshoot A_m is the maximum deviation of the step response output above its steady state value y_s.
- The peak time t_p is the time at which the maximum overshoot occurs.
- Settling time t_s is the time at which the step response input stays within some small percentage rage of the steady state value y_s. Typically, a percentage of 2% or 5% is chosen to determine the settling time.
- The rise time t_r is usually defined as the time required for the step response output to rise from 10% to 90% of the steady state value y_s.
- The delay time t_d is defined as the time required to reach 50% of the steady state value y_s.

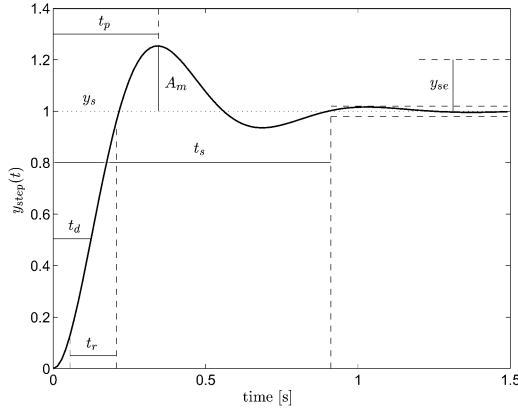



FIGURE 25.5 Parameters for step-response behavior: steady state value y_s, steady state error y_{se}, maximum overshoot A_m, peak time t_p, settling time t_s, rise time t_r, and delay time t_d.

Most of the above value can be obtained from an experimentally determined step response. In general, they cannot be obtained in an analytical form, except for low order models. For the second order model of the one mass system given in (25.3), some analytical results can be obtained. For a second order model of (25.3), the maximum overshoot A_m is determined by

$$A_m = 100e^{-\pi\beta}/\sqrt{1-\xi^2}, \quad \text{where } \xi = \frac{A}{\sqrt{\pi^2 + A^2}}, \quad A = \ln\left(\frac{100}{A_m}\right)$$

The peak time t_p can be computed by

$$t_p = \frac{\pi}{\omega_n\sqrt{1-\xi^2}}$$

whereas the delay time t_d can be approximated by

$$t_d \approx \frac{1+0.7\xi}{\omega_n}$$

As the maximum overshoot increases with a smaller damping coefficient β in the system, the maximum overshoot is often used to indicate the relative stability of the system.

Frequency Domain Parameters

With the frequency domain analysis of dynamic systems, specifications for the dynamic properties of a system can also be stated in the frequency domain. Frequency domain specifications in filter design often address ripple, bandwidth, roll-off, and phase lag parameters. Similar characteristics can also be specified for dynamic systems in case the model of the system is analyzed in the frequency domain. The most significant parameters have been summarized below and illustrated in Fig. 25.6.

- The bandwidth ω_b is a notion for the maximum frequency at which the output will track a sinusoidal input in a satisfactory manner. By convention, the bandwidth is defined as the frequency at which the output is attenuated −3 dB (0.707).
- The resonant frequency ω_r is the first frequency at which a significant resonance mode with low damping occurs. The resonance mode can, if uncontrolled, negatively influence the settling time of the dynamic system and plays an important role in characterization of performance.

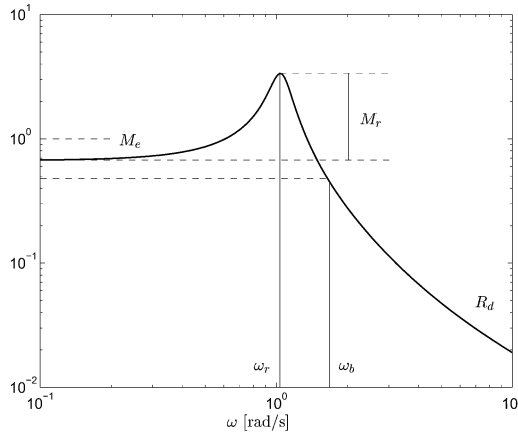



FIGURE 25.6 Parameters for frequency response behavior: bandwidth ω_b, resonance frequency ω_r, resonant peak M_r, steady state error M_e, and roll-off R_d.

- The resonant peak M_r is the height of a resonance mode. The resonant peak is a measure for the damping. As illustrated in Fig. 25.2 for a second order model, the resonance mode increases at lower damping coefficients.
- Steady state errors M_e can also be analyzed in the frequency response of a system. Using the final value theorem for continuous time systems

$$\lim_{t \to \infty} y(t) = y_s = \lim_{s \to 0} sy(s)$$

the presence of steady state errors can be inspected in the frequency domain by evaluation $|G(s)|$ at $s = i\omega = 0$ or for small values of the frequency vector ω. This can be seen as follows. As the Laplace transform $u_{\text{step}}(s)$ of a step input signal $u_{\text{step}}(t)$ is $u_{\text{step}}(s) = 1/s$,

$$\lim_{t \to \infty} y_{\text{step}}(t) = \lim_{s \to 0} sy_{\text{step}}(s) = \lim_{s \to 0} sG(s)\frac{1}{s} = \lim_{s \to 0} G(s)$$

By evaluating $|G(i\omega)|$ for small frequencies ω, the steady state behavior of $G(s)$ can be studied.

A similar result exist for discrete time systems, where the final value theorem reads as follows. If $u(z)$ converges for $|z| > 1$ and all poles of $(z - 1)u(z)$ are inside the unit circle, then

$$\lim_{k \to \infty} u(k) = \lim_{z \to 1} (z - 1)u(z)$$

Hence, for discrete time systems the steady state behavior of a transfer function $G(z)$ can be studied by evaluating $|G(e^{i\omega\Delta T})|$ for small frequencies ω.

- Roll-off R_d at high frequencies is defined as the negative slope of the frequency response at higher frequencies. The roll-off determines the performance of the dynamic system as high frequent disturbances can be amplified if a dynamic system does not have enough high frequent roll-off.

26

The Root Locus Method

Hitay Özbay
The Ohio State University

26.1 Introduction

The root locus technique is a graphical tool used in feedback control system analysis and design. It has been formally introduced to the engineering community by W. R. Evans [3,4], who received the Richard E. Bellman Control Heritage Award from the American Automatic Control Council in 1988 for this major contribution.

In order to discuss the root locus method, we must first review the basic definition of bounded input bounded output (BIBO) stability of the standard linear time invariant feedback system shown in Fig. 26.1, where the plant, and the controller, are represented by their transfer functions $P(s)$ and $C(s)$, respectively.[1] The plant, $P(s)$, includes the physical process to be controlled, as well as the actuator and the sensor dynamics.

The feedback system is said to be stable if none of the closed-loop transfer functions, from external inputs r and v to internal signals e and u, have any poles in the closed right half plane, $\overline{\mathbb{C}}_+ := \{s \in \mathbb{C} : \mathrm{Re}(s) \geq 0\}$. A necessary condition for feedback system stability is that the closed right half plane zeros of $P(s)$ (respectively $C(s)$) are distinct from the poles of $C(s)$ (respectively $P(s)$). When this condition holds, we say that there is no unstable pole–zero cancellation in taking the product $P(s)C(s) =: G(s)$, and then checking feedback system stability becomes equivalent to checking whether all the roots of

$$1 + G(s) = 0 \tag{26.1}$$

are in the open left half plane, $\mathbb{C}_- := \{s \in \mathbb{C} : \mathrm{Re}(s) < 0\}$. The roots of (26.1) are the closed-loop system poles. We would like to understand how the closed-loop system pole locations vary as functions of a real parameter of $G(s)$. More precisely, assume that $G(s)$ contains a parameter K, so that we use the notation

[1]Here we consider the continuous time case; there is essentially no difference between the continuous time case and the discrete time case, as far as the root locus construction is concerned. In the discrete time case the desired closed-loop pole locations are defined relative to the unit circle, whereas in the continuous time case desired pole locations are defined relative to the imaginary axis.

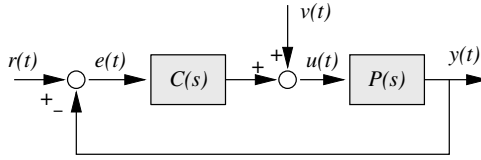



FIGURE 26.1 Standard unity feedback system.

$G(s) = G_K(s)$ to emphasize the dependence on K. The *root locus* is the plot of the roots of (26.1) on the complex plane, as the parameter K varies within a specified interval.

The most common example of the root locus problem deals with the uncertain (or adjustable) gain as the varying parameter: when $P(s)$ and $C(s)$ are fixed rational functions, except for a gain factor, $G(s)$ can be written as $G(s) = G_K(s) = KF(s)$, where K is the uncertain/adjustable gain, and

$$F(s) = \frac{N(s)}{D(s)} \quad \text{where} \quad \begin{aligned} N(s) &= \prod_{j=1}^{m}(s - z_j) \\ D(s) &= \prod_{i=1}^{n}(s - p_i), \end{aligned} \quad n \geq m \tag{26.2}$$

with $z_1,\ldots,z_m$, and $p_1,\ldots,p_n$ being the open-loop system zeros and poles. In this case, the closed-loop system poles are the roots of the characteristic equation

$$\chi(s) := D(s) + KN(s) = 0 \tag{26.3}$$

The *usual root locus* is obtained by plotting the roots $r_1(K),\ldots,r_n(K)$ of the characteristic polynomial $\chi(s)$ on the complex plane, as K varies from 0 to $+\infty$. The same plot for the negative values of K gives the *complementary root locus*. With the help of the root locus plot the designer identifies the admissible values of the parameter K leading to a set of closed-loop system poles that are in the desired region of the complex plane. There are several factors to be considered in defining the "desired region" of the complex plane in which all the roots $r_1(K),\ldots,r_n(K)$ should lie. Those are discussed briefly in the next section. Section 26.3 contains the root locus construction procedure, and design examples are presented in section 26.4.

The root locus can also be drawn with respect to a system parameter other than the gain. For example, the characteristic equation for the system $G(s) = G_\lambda(s)$, defined by

$$G_\lambda(s) = P(s)C(s), \quad P(s) = \frac{(1 - \lambda s)}{s(1 + \lambda s)}, \; C(s) = K_c\left(1 + \frac{1}{T_I s}\right)$$

can also be transformed into the form given in (26.3). Here K_c and T_I are given fixed PI (Proportional plus Integral) controller parameters, and $\lambda > 0$ is an uncertain plant parameter. Note that the phase of the plant is

$$\angle P(j\omega) = -\frac{\pi}{2} - 2\tan^{-1}(\lambda\omega)$$

so the parameter λ can be seen as the uncertain phase lag factor (for example, a small uncertain time delay in the plant can be modeled in this manner, see [9]). It is easy to see that the characteristic equation is

$$s^2(\lambda s + 1) + K_c(1 - \lambda s)\left(s + \frac{1}{T_I}\right) = 0$$

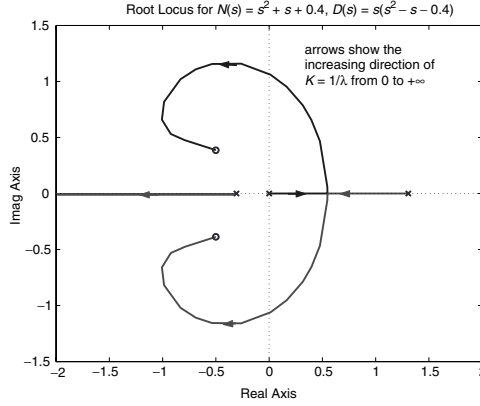



FIGURE 26.2 The root locus with respect to $K = 1/\lambda$.

and by rearranging the terms multiplying λ this equation can be transformed to

$$1 + \frac{1}{\lambda}\frac{(s^2 + K_c s + K_c/T_I)}{s(s^2 - K_c s - K_c/T_I)} = 0$$

By defining $K = \lambda^{-1}$, $N(s) = (s^2 + K_c s + K_c/T_I)$, and $D(s) = s(s^2 - K_c s - K_c/T_I)$, we see that the characteristic equation can be put in the form of (26.3). The root locus plot can now be obtained from the data $N(s)$ and $D(s)$ defined above; that shows how closed-loop system poles move as λ^{-1} varies from 0 to $+\infty$, for a given fixed set of controller parameters K_c and T_I. For the numerical example $K_c = 1$ and $T_I = 2.5$, the root locus is illustrated in Fig. 26.2.

The root locus construction procedure will be given in section 26.3. Most of the computations involved in each step of this procedure can be performed by hand calculations. Hence, an approximate graph representing the root locus can be drawn easily. There are also several software packages to generate the root locus automatically from the problem data $z_1,\ldots,z_m$, and $p_1,\ldots,p_n$.

If a numerical computation program is available for calculating the roots of a polynomial, we can also obtain the root locus with respect to a parameter which enters into the characteristic equation nonlinearly. To illustrate this point let us consider the following example: $G(s) = G_{\omega_o}(s)$ where

$$G_{\omega_o}(s) = P(s)C(s), \quad P(s) = \frac{(s - 0.1)}{(s^2 + 1.2\omega_o s + \omega_o^2)(s + 0.1)}, \quad C(s) = \frac{(s - 0.2)}{(s + 2)}$$

Here $\omega_o \geq 0$ is the uncertain plant parameter. Note that the characteristic equation

$$1 + \frac{\omega_o(1.2s + \omega_o)(s + 0.1)(s + 2)}{s^2(s + 0.1)(s + 2) + (s - 0.2)(s - 0.1)} = 0 \qquad (26.4)$$

cannot be expressed in the form of $D(s) + KN(s) = 0$ with a single parameter K. Nevertheless, for each ω_o we can numerically calculate the roots of (26.4) and plot them on the complex plane as ω_o varies within a range of interest. Figure 26.3 illustrates all the four branches, $r_1(K),\ldots,r_4(K)$, of the root locus for this system as ω_o increases from zero to infinity. The figure is obtained by computing the roots of (26.4) for a set of values of ω_o by using MATLAB.

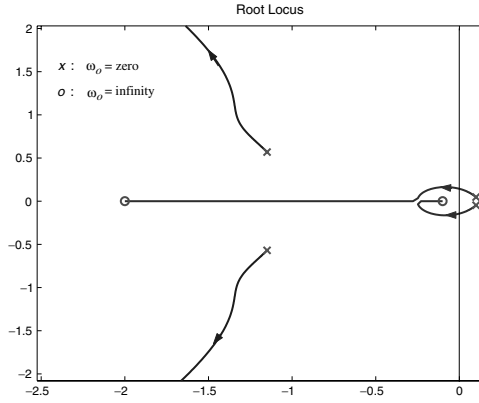



FIGURE 26.3 The root locus with respect to ω_o.

26.2 Desired Pole Locations

The performance of a feedback system depends heavily on the location of the closed-loop system poles $r_i(K) = 1,\ldots,n$. First of all, for stability we want $r_i(K) \in \mathbb{C}_-$ for all $i = 1,\ldots,n$. Clearly, having a pole "close" to the imaginary axis poses a danger, i.e., "small" perturbations in the plant might lead to an unstable feedback system. So the desired pole locations must be such that stability is preserved under such perturbations (or in the presence of uncertainties) in the plant. For second-order systems, we can define certain stability robustness measures in terms of the pole locations, which can be tied to the characteristics of the step response. For higher order systems, similar guidelines can be used by considering the dominant poles only.

In the standard feedback control system shown in Fig. 26.1, assume that the closed-loop transfer function from $r(t)$ to $y(t)$ is in the form

$$T(s) = \frac{\omega_o^2}{s^2 + 2\zeta\omega_o s + \omega_o^2}, \quad 0 < \zeta < 1, \quad \omega_o \in \mathbb{R}$$

and $r(t)$ is the unit step function. Then, the output is

$$y(t) = 1 - \frac{e^{-\zeta\omega_o t}}{\sqrt{1-\zeta^2}}\sin(\omega_d t + \theta), \qquad t \geq 0$$

where $\omega_d := \omega_o\sqrt{1-\zeta^2}$ and $\theta := \cos^{-1}(\zeta)$. For some typical values of ζ, the step response $y(t)$ is as shown in Fig. 26.4. The maximum *percent overshoot* is defined to be the quantity

$$\text{PO} := \frac{y_p - y_{ss}}{y_{ss}} \times 100\%$$

where y_p is the peak value. By simple calculations it can be seen that the peak value of $y(t)$ occurs at the time instant $t_p = \pi/\omega_d$, and

$$\text{PO} = e^{-\pi\zeta/\sqrt{1-\zeta^2}} \times 100\%$$

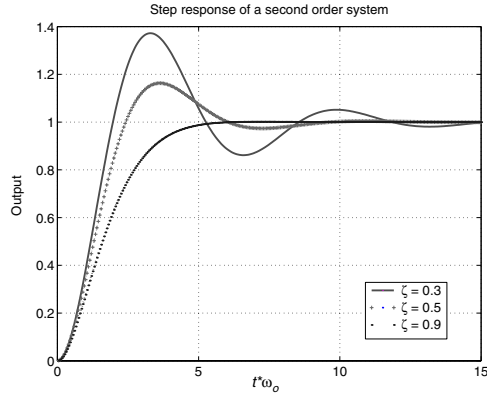



FIGURE 26.4 Step response of a second-order system.

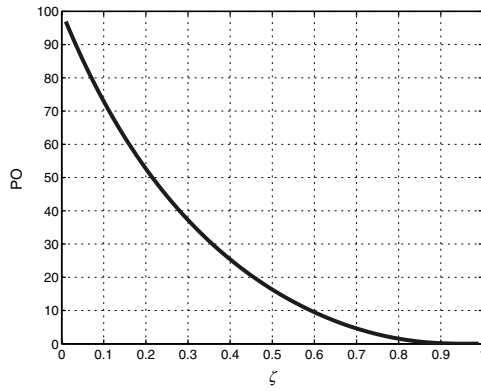



FIGURE 26.5 PO versus ζ.

Figure 26.5 shows PO versus ζ. The *settling time* is defined to be the smallest time instant t_s, after which the response $y(t)$ remains within 2% of its final value, i.e.,

$$t_s := \min\{t' : |y(t) - y_{ss}| \leq 0.02 y_{ss} \;\; \forall t \geq t'\}$$

Sometimes 1% or 5% is used in the definition of settling time instead of 2%; conceptually, there is no difference. For the second-order system response, we have

$$t_s \approx \frac{4}{\zeta \omega_o}$$

So, in order to have a fast settling response, the product $\zeta\omega_o$ should be large.

The closed-loop system poles are

$$r_{1,2} = -\zeta\omega_o \pm j\omega_o\sqrt{1-\zeta^2}$$

Therefore, once the maximum allowable settling time and PO are specified, we can define the region of desired pole locations by determining the minimum allowable ζ and $\zeta\omega_o$. For example, let the desired PO and t_s be bounded by

$$\text{PO} \leq 10\% \quad \text{and} \quad t_s \leq 8\,\text{s}$$

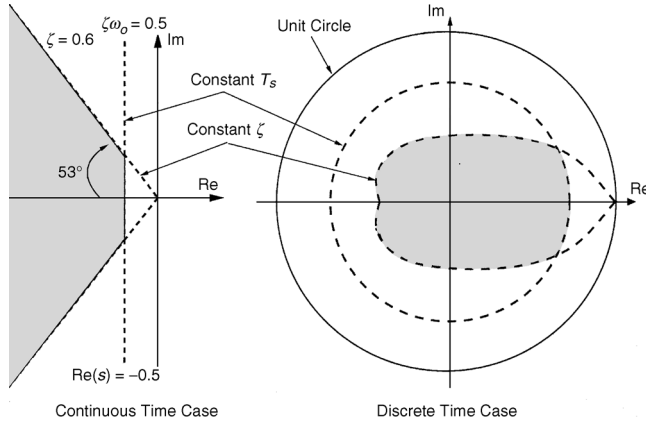



FIGURE 26.6 Region of the desired closed-loop poles.

The PO requirement implies that $\zeta \geq 0.6$, equivalently $\theta \leq 53°$ (recall that $\cos(\theta) = \zeta$). The settling time requirement is satisfied if and only if $\text{Re}(r_{1,2}) \leq -0.5$. Then, the region of desired closed-loop poles is the shaded area shown in Fig. 26.6. The same figure also illustrates the region of desired closed-loop poles for similar design requirements in the discrete time case.

If the order of the closed-loop transfer function $T(s)$ is higher than two, then, depending on the location of its poles and zeros, it may be possible to approximate the closed-loop step response by the response of a second-order system. For example, consider the third-order system

$$T(s) = \frac{\omega_o^2}{(s^2 + 2\zeta\omega_o s + \omega_o^2)(1 + s/r)} \quad \text{where } r \gg \zeta\omega_o$$

The transient response contains a term e^{-rt}. Compared with the envelope $e^{-\zeta\omega_o t}$ of the sinusoidal term, e^{-rt} decays very fast, and the overall response is similar to the response of a second-order system. Hence, the effect of the third pole $r_3 = -r$ is negligible.

Consider another example,

$$T(s) = \frac{\omega_o^2[1 + s/(r+\epsilon)]}{(s^2 + 2\zeta\omega_o s + \omega_o^2)(1 + s/r)} \quad \text{where } 0 < \epsilon \ll r$$

In this case, although r does not need to be much larger than $\zeta\omega_o$, the zero at $-(r + \epsilon)$ cancels the effect of the pole at $-r$. To see this, consider the partial fraction expansion of $Y(s) = T(s)R(s)$ with $R(s) = 1/s$:

$$Y(s) = \frac{A_0}{s} + \frac{A_1}{s - r_1} + \frac{A_2}{s - r_2} + \frac{A_3}{s + r}$$

where $A_0 = 1$ and

$$A_3 = \lim_{s \to -r}(s + r)Y(s) = \frac{\omega_o^2}{2\zeta\omega_o r - (\omega_o^2 + r^2)}\left(\frac{\epsilon}{r + \epsilon}\right)$$

Since $|A_3| \to 0$ as $\epsilon \to 0$, the term $A_3 e^{-rt}$ is negligible in $y(t)$.

In summary, if there is an approximate pole–zero cancellation in the left half plane, then this pole–zero pair can be taken out of the transfer function $T(s)$ to determine PO and t_s. Also, the poles closest to the

imaginary axis dominate the transient response of $y(t)$. To generalize this observation, let $r_1,\ldots,r_n$ be the poles of $T(s)$, such that $\mathrm{Re}(r_k) \ll \mathrm{Re}(r_2) = \mathrm{Re}(r_1) < 0$, for all $k \geq 3$. Then, the pair of complex conjugate poles $r_{1,2}$ are called the *dominant poles*. We have seen that the desired transient response properties, e.g., PO and t_s, can be translated into requirements on the location of the dominant poles.

26.3 Root Locus Construction

As mentioned above, the root locus primarily deals with finding the roots of a characteristic polynomial that is an affine function of a single parameter, K,

$$\chi(s) = D(s) + KN(s) \tag{26.5}$$

where $D(s)$ and $N(s)$ are fixed monic polynomials (i.e., coefficient of the highest power is normalized to 1). If N and/or D are not monic, the highest coefficient(s) can be absorbed into K.

Root Locus Rules

Recall that the usual root locus shows the locations of the closed-loop system poles as K varies from 0 to $+\infty$. The roots of $D(s)$, $p_1,\ldots,p_n$, are the poles, and the roots of $N(s)$, $z_1,\ldots,z_m$, are the zeros, of the open-loop system, $G(s) = KF(s)$. Since $P(s)$ and $C(s)$ are proper, $G(s)$ is proper, and hence $n \geq m$. So the degree of the polynomial $\chi(s)$ is n and it has exactly n roots.

Let the closed-loop system poles, i.e., roots of $\chi(s)$, be denoted by $r_1(K),\ldots,r_n(K)$. Note that these are functions of K; whenever the dependence on K is clear, they are simply written as $r_1,\ldots,r_n$. The points in $\mathbb{C}$ that satisfy (26.5) for some $K > 0$ are on the root locus. Clearly, a point $r \in \mathbb{C}$ is on the root locus if and only if

$$K = -\frac{1}{F(r)} \tag{26.6}$$

The condition (26.6) can be separated into two parts:

$$|K| = -\frac{1}{|F(r)|} \tag{26.7}$$

$$\angle K = 0^\circ = -(2\ell+1)\times 180^\circ - \angle F(r), \quad \ell = 0, \pm 1, \pm 2, \ldots. \tag{26.8}$$

The phase rule (26.8) determines the points in $\mathbb{C}$ that are on the root locus. The magnitude rule (26.7) determines the gain $K > 0$ for which the root locus is at a given point r. By using the definition of $F(s)$, (26.8) can be rewritten as

$$(2\ell+1)\times 180^\circ = \sum_{i=1}^{n}\angle(r-p_i) - \sum_{j=1}^{m}\angle(r-z_j) \tag{26.9}$$

Similarly, (26.7) is equivalent to

$$K = \frac{\prod_{i=1}^{n}|r-p_i|}{\prod_{j=1}^{m}|r-z_j|} \tag{26.10}$$

Root Locus Construction

There are several software packages available for generating the root locus automatically for a given $F = N/D$. In particular, the related MATLAB commands are `rlocus` and `rlocfind`. In many cases, approximate root locus can be drawn by hand using the rules given below. These rules are determined from the basic definitions (26.5), (26.7), and (26.8).

1. The root locus has n branches: $r_1(K),\ldots,r_n(K)$.
2. Each branch starts $(K \cong 0)$ at a pole p_i and ends (as $K \to \infty$) at a zero z_j, or converges to an asymptote, $Me^{j\alpha\ell}$, where $M \to \infty$ and

$$\alpha_\ell = \frac{2\ell + 1}{n - m} \times 180°, \quad \ell = 0,\ldots,(n - m - 1)$$

3. There are $(n - m)$ asymptotes with angles α_ℓ. The center of the asymptotes (i.e., their intersection point on the real axis) is

$$\sigma_a = \frac{\sum_{i=1}^{n} p_i - \sum_{j=1}^{m} z_j}{n - m}$$

4. A point $x \in \mathbb{R}$ is on the root locus if and only if the total number of poles p_i's and zeros z_j's to the right of x (i.e., total number of p_i's with $\text{Re}(p_i) > x$ plus total number of z_j's with $\text{Re}(z_j) > x$) is odd. Since $F(s)$ is a rational function with real coefficients, poles and zeros appear in complex conjugates, so when counting the number of poles and zeros to the right of a point $x \in \mathbb{R}$ we just need to consider the poles and zeros on the real axis.
5. The values of K for which the root locus crosses the imaginary axis can be determined from the Routh–Hurwitz stability test. Alternatively, we can set $s = j\omega$ in (26.5) and solve for real ω and K satisfying

$$D(j\omega) + KN(j\omega) = 0$$

   Note that there are two equations here, one for the real part and one for the imaginary part.
6. The break points (intersection of two branches on the real axis) are feasible solutions (satisfying rule 4) of

$$\frac{d}{ds}F(s) = 0 \tag{26.11}$$

7. Angles of departure $(K \cong 0)$ from a complex pole, or arrival $(K \to +\infty)$ to a complex zero, can be determined from the phase rule. See example below.

   Let us now follow the above rules step by step to construct the root locus for

$$F(s) = \frac{(s + 3)}{(s - 1)(s + 5)(s + 4 + j2)(s + 4 - j2)}$$

First, enumerate the poles and zeros as $p_1 = -4 + j2$, $p_2 = -4 - j2$, $p_3 = -5$, $p_4 = 1$, $z_1 = -3$. So, $n = 4$ and $m = 1$.

1. The root locus has four branches.
2. Three branches converge to the asymptotes whose angles are 60°, 180°, and –60°, and one branch converges to $z_1 = -3$.

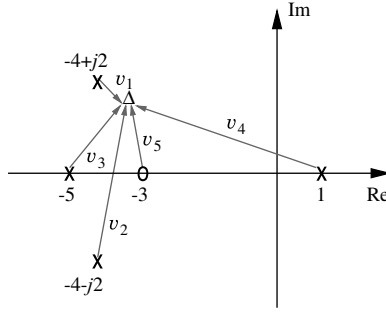



FIGURE 26.7 Angle of departure from $-4 + j2$.

3. The center of the asymptotes is $\sigma = (-12 + 3)/3 = -3$.
4. The intervals $(-\infty, -5]$ and $[-3, 1]$ are on the root locus.
5. The imaginary axis crossings are the feasible roots of

$$(\omega^4 - j12\omega^3 - 47\omega^2 + j40\omega - 100) + K(j\omega + 3) = 0 \tag{26.12}$$

for real ω and K. Real and imaginary parts of (26.12) are

$$\omega^4 - 47\omega^2 - 100 + 3K = 0$$
$$j\omega(-12\omega^2 + 40 + K) = 0$$

They lead to two feasible pairs of solutions ($K = 100/3$, $\omega = 0$) and ($K = 215.83$, $\omega = \pm 4.62$).
6. Break points are the feasible solutions of

$$3s^4 + 36s^3 + 155s^2 + 282s + 220 = 0$$

Since the roots of this equation are $-4.55 \pm j1.11$ and $-1.45 \pm j1.11$, there is no solution on the real axis, hence no break points.
7. To determine the angle of departure from the complex pole $p_1 = -4 + j2$, let Δ represent a point on the root locus near the complex pole p_1, and define v_i, $i = 1,\ldots,5$, to be the vectors drawn from p_i, for $i = 1,\ldots,4$, and from z_1 for $i = 5$, as shown in Fig. 26.7. Let $\theta_1,\ldots,\theta_5$ be the angles of $v_1,\ldots,v_5$. The phase rule implies

$$(\theta_1 + \theta_2 + \theta_3 + \theta_4) - \theta_5 = \pm 180° \tag{26.13}$$

As Δ approaches p_1, θ_1 becomes the angle of departure and the other θ_i's can be approximated by the angles of the vectors drawn from the other poles, and from the zero, to the pole p_1. Thus θ_1 can be solved from (26.13), where $\theta_2 \approx 90°$, $\theta_3 \approx \tan^{-1}(2)$, $\theta_4 \approx 180° - \tan^{-1}\left(\frac{2}{5}\right)$, and $\theta_5 \approx 90° + \tan^{-1}\left(\frac{1}{2}\right)$. That yields $\theta_1 \approx -15°$.

The exact root locus for this example is shown in Fig. 26.8. From the results of item 5 above, and the shape of the root locus, it is concluded that the feedback system is stable if

$$33.33 < K < 215.83$$

i.e., by simply adjusting the gain of the controller, the system can be made stable. In some situations we need to use a dynamic controller to satisfy all the design requirements.

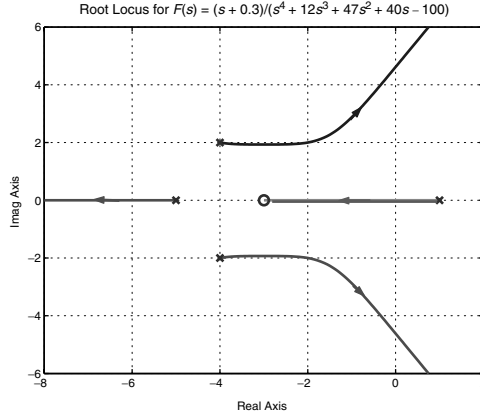



FIGURE 26.8 Root locus for $F(s) = \dfrac{(s+3)}{(s-1)(s+5)(s+4+j2)(s+4-j2)}$.

Design Examples

Example 1

Consider the standard feedback system with a plant

$$P(s) = \frac{1}{0.72}\frac{1}{(s+1)(s+2)}$$

and design a controller such that

- the feedback system is stable,
- $\mathrm{PO} \le 10\%$, $t_s \le 4$ s, and steady state error is zero when $r(t)$ is unit step,
- steady state error is as small as possible when $r(t)$ is unit ramp.

It is clear that the second design goal cannot be achieved by a simple proportional controller. To satisfy this condition, the controller must have a pole at $s = 0$, i.e., it must have integral action. If we try an integral control of the form $C(s) = K_c/s$, with $K_c > 0$, then the root locus has three branches, the interval $[-1, 0]$ is on the root locus; three asymptotes have angles $\{60°, 180°, -60°\}$ with a center at $\sigma_a = -1$; and there is only one break point at $-1 + \frac{1}{\sqrt{3}}$, see Fig. 26.9. From the location of the break point, center, and angles of the asymptotes, it can be deduced that two branches (one starting at $p_1 = -1$, and the other one starting at $p_3 = 0$) always remain to the right of p_1. On the other hand, the settling time condition implies that the real parts of the dominant closed-loop system poles must be less than or equal to -1. So, a simple integral control does not do the job. Now try a PI controller of the form

$$C(s) = K_c\left(\frac{s - z_c}{s}\right), \quad K_c > 0$$

In this case, we can select $z_c = -1$ to cancel the pole at $p_1 = -1$ and the system effectively becomes a second-order system. The root locus for $F(s) = 1/s(s + 2)$ has two branches and two asymptotes, with center $\sigma_a = -1$ and angles $\{90°, -90°\}$; the break point is also at -1. The branches leave -2 and 0, and go toward each other, meet at -1, and tend to infinity along the line $\mathrm{Re}(s) = -1$. Indeed, the closed-loop system poles are

$$r_{1,2} = -1 \pm \sqrt{1-K}, \quad \text{where } K = K_c/0.72$$

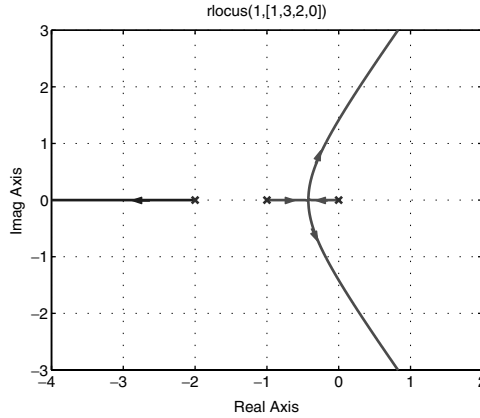



FIGURE 26.9 Root locus for Example 1.

The steady state error, when $r(t)$ is unit ramp, is $2/K$. So K needs to be as large as possible to meet the third design condition. Clearly, $\mathrm{Re}(r_{1,2}) = -1$ for all $K \geq 1$, which satisfies the settling time requirement. The percent overshoot is less than 10% if ζ of the roots $r_{1,2}$ is greater than 0.6. A simple algebra shows that $\zeta = 1/\sqrt{K}$, hence the design conditions are met if $K = 1/0.36$, i.e. $K_c = 2$. Thus a PI controller that solves the design problem is

$$C(s) = 2\left(\frac{s+1}{s}\right)$$

The controller cancels a stable pole (at $s = -1$) of the plant. If there is a slight uncertainty in this pole location, perfect cancellation will not occur and the system will be third-order with the third pole at $r_3 \cong -1$. Since the zero at $z_o = -1$ will approximately cancel the effect of this pole, the response of this system will be close to the response of a second-order system. However, we must be careful if the pole–zero cancellations are near the imaginary axis because in this case small perturbations in the pole location might lead to large variations in the feedback system response, as illustrated with the next example.

Example 2

A flexible structure with lightly damped poles has transfer function in the form

$$P(s) = \frac{\omega_1^2}{s^2(s^2 + 2\zeta\omega_1 s + \omega_1^2)}$$

By using the root locus, we can see that the controller

$$C(s) = K_c \frac{(s^2 + 2\zeta\omega_1 s + \omega_1^2)(s + 0.4)}{(s + r)^2(s + 4)}$$

stabilizes the feedback system for sufficiently large r and an appropriate choice of K_c. For example, let $\omega_1 = 2$, $\zeta = 0.1$, and $r = 10$. Then the root locus of $F(s) = P(s)C(s)/K$, where $K = K_c\omega_1^2$, is as shown in Fig. 26.10. For $K = 600$, the closed-loop system poles are

$$\{-10.78 \pm j2.57, -0.94 \pm j1.61, -0.2 \pm j1.99, -0.56\}$$

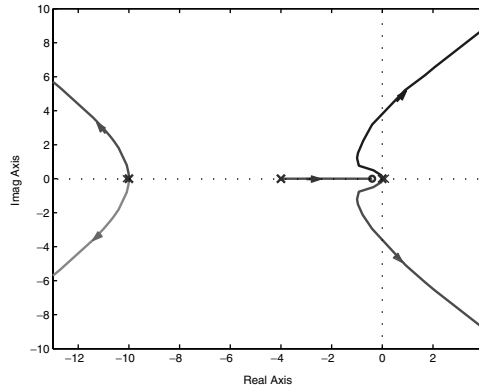



FIGURE 26.10 Root locus for Example 2(a).

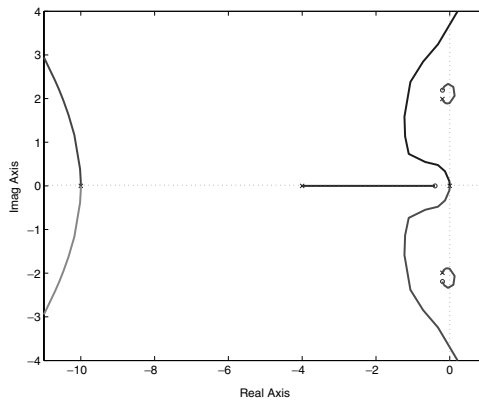



FIGURE 26.11 Root locus for Example 2(b).

Since the poles $-0.2 \pm j1.99$ are canceled by a pair of zeros at the same point in the closed-loop system transfer function $T = G(1 + G)^{-1}$, the dominant poles are at -0.56 and $-0.94 \pm j1.61$ (they have relatively large negative real parts and the damping ratio is about 0.5).

Now, suppose that this controller is fixed and the complex poles of the plant are slightly modified by taking $\zeta = 0.09$ and $\omega_1 = 2.2$. The root locus corresponding to this system is as shown in Fig. 26.11. Since lightly damped complex poles are not perfectly canceled, there are two more branches near the imaginary axis. Moreover, for the same value of $K = 600$, the closed-loop system poles are

$$\{-10.78 \pm j2.57, -1.21 \pm j1.86, 0.05 \pm j1.93, -0.51\}$$

In this case, the feedback system is unstable.

Example 3

One of the most important examples of mechatronic systems is the DC motor. An approximate transfer function of a DC motor [8, pp. 141–143] is in the form

$$P_m(s) = \frac{K_m}{s(s + 1/\tau_m)}, \quad \tau_m > 0$$

Also note that if τ_m is large, then $P_m(s) \approx P_b(s)$, where $P_b(s) = K_b/s^2$ is the transfer function of a rigid beam. In this example, the general class of plants $P_m(s)$ will be considered. Assuming that $p_m = -1/\tau_m$ and K_m are given, a first-order controller

$$C(s) = K_c\left(\frac{s - z_c}{s - p_c}\right) \tag{26.14}$$

will be designed. The aim is to place the closed-loop system poles far from the Im-axis. Since the order of $F(s) = P_m(s)C(s)/K_mK_c$ is three, the root locus has three branches. Suppose the desired closed-loop poles are given as p_1, p_2, and p_3. Then, the pole placement problem amounts to finding $\{K_c, z_c, p_c\}$ such that the characteristic equation is

$$\begin{aligned}\chi(s) &= (s-p_1)(s-p_2)(s-p_3)\\ &= s^3 - (p_1+p_2+p_3)s^2 + (p_1p_2+p_1p_3+p_2p_3)s - p_1p_2p_3\end{aligned}$$

But the actual characteristic equation, in terms of the unknown controller parameters, is

$$\begin{aligned}\chi(s) &= s(s-p_m)(s-p_c) + k(s-z_c)\\ &= s^3 - (p_m+p_c)s^2 + (p_mp_c+K)s - Kz_c\end{aligned}$$

where $K := K_mK_c$. Equating the coefficients of the desired $\chi(s)$ to the coefficients of the actual $\chi(s)$, three equations in three unknowns are obtained:

$$\begin{aligned}p_m + p_c &= p_1 + p_2 + p_3\\ p_mp_c + K &= p_1p_2 + p_1p_3 + p_2p_3\\ Kz_c &= p_1p_2p_3\end{aligned}$$

From the first equation p_c is determined, then K is obtained from the second equation, and finally z_c is computed from the third equation.

For different numerical values of p_m, p_1, p_2, and p_3 the shape of the root locus is different. Below are some examples, with the corresponding root loci shown in Figs. 26.12–26.14.

(a) $p_m = -0.05$, $p_1 = p_2 = p_3 = -2 \Rightarrow$

$$K = 11.70, \quad p_c = -5.95, \quad z_c = -0.68$$

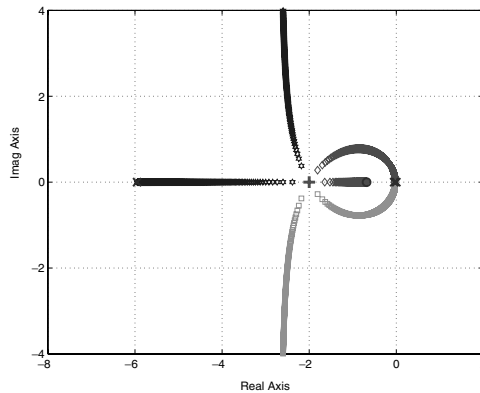



FIGURE 26.12 Root locus for Example 3(a).

FIGURE 26.13 Root locus for Example 3(b).

FIGURE 26.14 Root locus for Example 3(c).

(b) $p_m = -0.5,\ p_1 = -1,\ p_2 = -2,\ p_3 = -3 \Rightarrow$

$$K = 8.25, \quad p_c = -5.50, \quad z_c = -0.73$$

(c) $p_m = -5,\ p_1 = -11,\ p_2 = -4 + j1,\ p_3 = -4 - j1 \Rightarrow$

$$K = 35, \quad p_c = -14, \quad z_c = -5.343$$

Example 4

Consider the open-loop transfer function

$$P(s)C(s) = K_c \frac{(s^2 - 3s + 3)(s - z_c)}{s(s^2 + 3s + 3)(s - p_c)}$$

where K_c is the controller gain to be adjusted, and z_c and p_c are the controller zero and pole, respectively. Observe that the root locus has four branches except for the non-generic case $z_c = p_c$. Let the desired dominant closed-loop poles be $r_{1,2} = -0.4$. The steady state error for unit ramp reference input is

$$e_{ss} = \frac{p_c}{K_c z_c}$$

Accordingly, we want to make the ratio $K_c z_c / p_c$ as large as possible.

The characteristic equation is

$$\chi(s) = s(s^2 + 3s + 3)(s - p_c) + K_c(s^2 - 3s + 3)(s - z_c)$$

and it is desired to be in the form

$$\chi(s) = (s + 0.4)^2 (s - r_3)(s - r_4)$$

for some $r_{3,4}$ with $\mathrm{Re}(r_{3,4}) < 0$, which implies that

$$\chi(s)|_{s=-0.4} = 0, \quad \frac{d}{ds}\chi(s)|_{s=-0.4} = 0 \tag{26.15}$$

Conditions (26.15) give two equations:

$$0.784(0.4 + p_c) - 4.36K_c(0.4 + z_c) = 0$$
$$4.36K_c - 0.784 - 1.08(0.4 + p_c) + 3.8K_c(0.4 + z_c) = 0$$

from which z_c and p_c can be solved in terms of K_c. Then, by simple substitutions, the ratio to be maximized, $K_c z_c / p_c$, can be reduced to

$$\frac{K_c z_z}{p_c} = \frac{3.4776K_c - 0.784}{24.2469K_c - 3.4776}$$

The maximizing value of K_c is 0.1297; it leads to $p_c = -0.9508$ and $z_c = -1.1637$. For this controller, the feedback system poles are

$$\{-1.64 + j0.37, \ -1.64 - j0.37, \ -0.40, \ -0.40\}$$

The root locus is shown in Fig. 26.15.

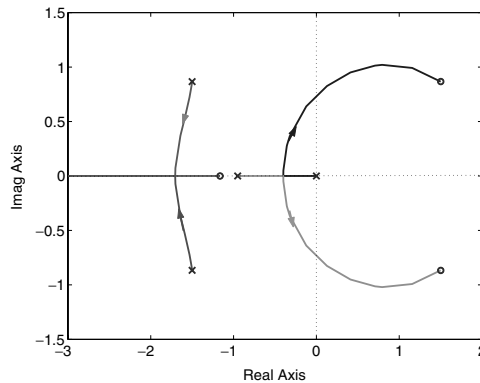



FIGURE 26.15 Root locus for Example 4.

26.4 Complementary Root Locus

In the previous section, the root locus parameter K was assumed to be positive and the phase and magnitude rules were established based on this assumption. There are some situations in which controller gain can be negative as well. Therefore, the complete picture is obtained by drawing the usual root locus (for $K > 0$) and the complementary root locus (for $K < 0$). The complementary root locus rules are

$$\ell \times 360° = \sum_{i=1}^{n} \angle(r - p_i) - \sum_{j=1}^{m} \angle(r - z_j), \quad \ell = 0, \pm 1, \pm 2, \ldots \tag{26.16}$$

$$|K| = \frac{\prod_{i=1}^{n} |r - p_i|}{\prod_{j=1}^{m} |r - z_j|} \tag{26.17}$$

Since the phase rule (26.16) is the 180° shifted version of (26.9), the complementary root locus is obtained by simple modifications in the root locus construction rules. In particular, the number of asymptotes and their center are the same, but their angles α_ℓ's are given by

$$\alpha_\ell = \frac{2\ell}{(n - m)} \times 180°, \quad \ell = 0, \ldots, (n - m - 1)$$

Also, an interval on the real axis is on the complementary root locus if and only if it is not on the usual root locus.

Example 3 (revisited)

In the Example 3 given above, if the problem data is modified to $p_m = -5$, $p_1 = -20$, and $p_{2,3} = -2 \pm j$, then the controller parameters become

$$K = -10, \quad p_c = -19, \quad z_c = 10$$

Note that the gain is negative. The roots of the characteristic equation as K varies between 0 and $-\infty$ form the complementary root locus; see Fig. 26.16.

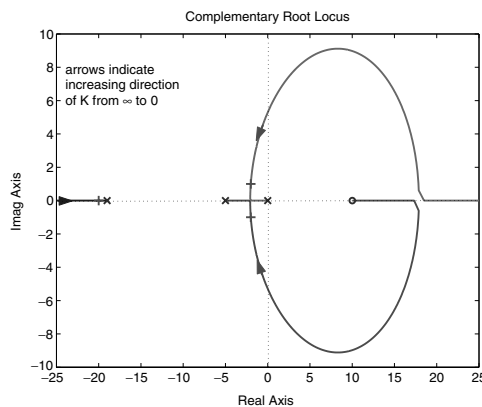



FIGURE 26.16 Complementary root locus for Example 3.

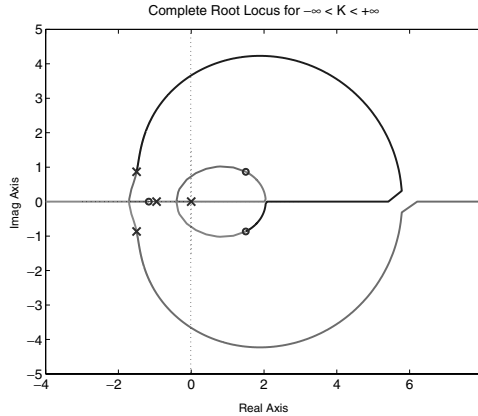



FIGURE 26.17 Complementary and usual root loci for Example 4.

Example 4 (revisited)

In this example, if K increases from $-\infty$ to $+\infty$, the closed-loop system poles move along the complementary root locus, and then the usual root locus, as illustrated in Fig. 26.17.

26.5 Root Locus for Systems with Time Delays

The standard feedback control system considered in this section is shown in Fig. 26.18, where the controller C and plant P are in the form

$$C(s) = \frac{N_c(s)}{D_c(s)}$$

and

$$P(s) = e^{-hs}P_0(s) \quad \text{where } P_0(s) = \frac{N_p(s)}{D_p(s)}$$

with (N_c, D_c) and (N_p, D_p) being coprime pairs of polynomials with real coefficients.[2] The term e^{-hs} is the transfer function of a pure delay element (in Fig. 26.18 the plant input is delayed by h seconds). In general, time delays enter into the plant model when there is

- a sensor (or actuator) processing delay, and/or
- a software delay in the controller, and/or
- a transport delay in the process.

In this case the open-loop transfer function is

$$G(s) = G_h(s) = e^{-hs}G_0(s)$$

where $G_0(s) = P_0(s)C(s)$ corresponds to the no delay case, $h = 0$.

Note that magnitude and phase of $G(j\omega)$ are determined from the identities

$$|G(j\omega)| = |G_0(j\omega)| \tag{26.18}$$

$$\angle G(j\omega) = -h\omega + \angle G_0(j\omega) \tag{26.19}$$

[2]A pair of polynomials is said to be coprime pair if they do not have common roots.

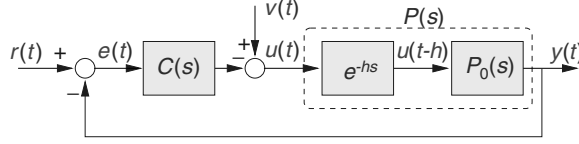



FIGURE 26.18 Feedback system a with time delay.

Stability of Delay Systems

Stability of the feedback system shown in Fig. 26.18 is equivalent to having all the roots of

$$\chi(s) \;=\; D(s) + e^{-hs}N(s) \tag{26.20}$$

in the open left half plane, $\mathbb{C}_-$, where $D(s) = D_c(s)D_p(s)$ and $N(s) = N_c(s)N_p(s)$. We assume that there is no unstable pole–zero cancellation in taking the product $P_0(s)C(s)$, and that $deg(D) > deg(N)$ (here N and D need not be monic polynomials). Strictly speaking, $\chi(s)$ is not a polynomial because it is a transcendental function of s. The functions of the form (26.20) belong to a special class of functions called *quasi-polynomials*. The closed-loop system poles are the roots of (26.20).

Following are known facts (see [1,10]):

(i) If r_k is a root of (20), then so is $\bar{r}_k$ (i.e., roots appear in complex conjugate pairs as usual).
(ii) There are infinitely many poles $r_k \in \mathbb{C}$, $k = 1, 2, \ldots$,satisfying $\chi(r_k) = 0$.
(iii) And r_k's can be enumerated in such a way that $\mathrm{Re}(r_k + 1) \leq \mathrm{Re}(r_k)$; moreover, $\mathrm{Re}(r_k) \to -\infty$ as $k \to \infty$.

Example

If $G_h(s) = e^{-hs}/s$, then the closed-loop system poles r_k, for $k = 1, 2, \ldots$, are the roots of

$$1 + \frac{e^{-h\sigma_k}e^{-jh\omega_k}}{\sigma_k + j\omega_k}e^{\pm j2k\pi} = 0 \tag{26.21}$$

where $r_k = \sigma_k + j\omega_k$ for some $\sigma_k, \omega_k \in \mathbb{R}$. Note that $e^{\pm j2k\pi} = 1$ for all $k = 1, 2, \ldots$. Equation (26.1) is equivalent to the following set of equations:

$$e^{-h\sigma_k} \;=\; |\sigma_k + j\omega_k| \tag{26.22}$$

$$\pm(2k-1)\pi \;=\; h\omega_k + \angle(\sigma_k + j\omega_k), \quad k = 1,2,\ldots \tag{26.23}$$

It is quite interesting that for $h = 0$ there is only one root $r = -1$, but even for infinitesimally small $h > 0$ there are infinitely many roots. From the magnitude condition (26.22), it can be shown that

$$\sigma_k \geq 0 \Rightarrow |\omega_k| \leq 1 \tag{26.24}$$

Also, for $\sigma_k \geq 0$, the phase $\angle(\sigma_k + j\omega_k)$ is between $-\pi/2$ and $+\pi/2$, therefore (26.23) leads to

$$\sigma_k \geq 0 \Rightarrow h|\omega_k| \geq \frac{\pi}{2} \tag{26.25}$$

By combining (26.24) and (26.25), it can be proven that the feedback system has no roots in the closed right half plane when $h < \pi/2$. Furthermore, the system is unstable if $h \geq \pi/2$. In particular, for $h = \pi/2$ there are two roots on the imaginary axis, at $\pm j1$. It is also easy to show that, for any $h > 0$ as $k \to \infty$, the roots converge to

$$r_k \to \frac{1}{h}\left[-\ln\left(\frac{2k\pi}{h}\right) \pm j2k\pi\right]$$

As $h \to 0$, the magnitude of the roots converge to ∞.

As illustrated by the above example, property (iii) implies that for any given real number σ there are only finitely many r_k's in the region of the complex plane

$$\mathbb{C}_\sigma := \{s \in \mathbb{C}: \mathrm{Re}(s) \geq \sigma\}$$

In particular, with $\sigma = 0$, this means that the quasi-polynomial $\chi(s)$ can have only finitely many roots in the right half plane. Since the effect of the closed-loop system poles that have very large negative real parts is negligible (as far as closed-loop systems' input–output behavior is concerned), only finitely many "dominant" roots r_k, for $k = 1,\ldots,m$, should be computed for all practical purposes.

Dominant Roots of a Quasi-Polynomial

Now we discuss the following problem: given $N(s)$, $D(s)$, and $h \geq 0$, find the dominant roots of the quasi-polynomial

$$\chi(s) = D(s) + e^{-hs}N(s)$$

For each fixed $h > 0$, it can be shown that there exists $\sigma_{\max}$ such that $\chi(s)$ has no roots in the region $\mathbb{C}_{\sigma_{\max}}$, see [11] for a simple algorithm to estimate $\sigma_{\max}$, based on Nyquist criterion. Given $h > 0$ and a region of the complex plane defined by $\sigma_{\min} \leq \mathrm{Re}(s) \leq \sigma_{\max}$, the problem is to find the roots of $\chi(s)$ in this region.

Clearly, a point $r = \sigma + j\omega$ in $\mathbb{C}$ is a root of $\chi(s)$ if and only if

$$D(\sigma + j\omega) = -e^{-h\sigma}e^{-jh\omega}N(\sigma + j\omega)$$

Taking the magnitude square of both sides of the above equation, $\chi(r) = 0$ implies

$$A_\sigma(x) := D(\sigma + x)D(\sigma - x) - e^{-2h\sigma}N(\sigma + x)N(\sigma - x) = 0$$

where $x = j\omega$. The term $D(\sigma + x)$ stands for the function $D(s)$ evaluated at $\sigma + x$. The other terms of $A_\sigma(x)$ are calculated similarly. For each fixed σ, the function $A_\sigma(x)$ is a polynomial in the variable x. By symmetry, if x is a zero of $A_\sigma(\cdot)$, then $(-x)$ is also a zero.

If $A_\sigma(x)$ has a root x_ℓ whose real part is zero, set $r_\ell = \sigma + x_\ell$. Next, evaluate the magnitude of $\chi(r_\ell)$; if it is zero, then n_ℓ is a root of $\chi(s)$. Conversely, if $A_\sigma(x)$ has no root on the imaginary axis, then $\chi(s)$ cannot have a root whose real part is the fixed value of σ from which $A_\sigma(\cdot)$ is constructed.

Algorithm

Given $N(s)$, $D(s)$, h, $\sigma_{\min}$, and $\sigma_{\max}$:

Step 1. Pick σ values $\sigma_1,\ldots,\sigma_M$ between $\sigma_{\min}$ and $\sigma_{\max}$ such that $\sigma_{\min} = \sigma_1$, $\sigma_i < \sigma_{i+1}$, and $\sigma_M = \sigma_{\max}$. For each σ_i perform the following.

Step 2. Construct the polynomial $A_i(x)$ according to

$$A_i(x) := D(\sigma_i + x)D(\sigma_i - x) - e^{-2h\sigma_i}N(\sigma_i + x)N(\sigma_i - x)$$

Step 3. For each imaginary axis roots x_ℓ of A_i, perform the following test:

Check if $|\chi(\sigma_i + x_\ell)| = 0$; if yes, then $r = \sigma_i + x_\ell$ is a root of $\chi(s)$; if not, discard x_ℓ.

Step 4. If $i = M$, stop; else increase i by 1 and go to Step 2.

Example

We will find the dominant roots of

$$1 + \frac{e^{-hs}}{s} = 0 \tag{26.26}$$

for a set of critical values of h. Recall that (26.26) has a pair of roots $\pm j1$ when $h = \pi/2 = 1.57$. Moreover, dominant roots of (26.26) are in the right half plane if $h > 1.57$, and they are in the left half plane if $h < 1.57$. So, it is expected that for $h \in (1.2, 2.0)$ the dominant roots are near the imaginary axis. Take $\sigma_{min} = -0.5$ and $\sigma_{max} = 0.5$, with $M = 400$ linearly spaced σ_i's between them. In this case

$$A_i(x) = \sigma_i^2 - e^{-2h\sigma_i} - x^2$$

Whenever $e^{-2h\sigma_i} \geq \sigma_i^2$, $A_i(x)$ has two roots:

$$x_\ell = \pm j\sqrt{e^{-2h\sigma_i} - \sigma_i^2}, \quad \ell = 1, 2$$

For each fixed σ_i satisfying this condition, let $r_\ell = \sigma_i + x_\ell$ (note that x_ℓ is a function of σ_i, so r_ℓ is a function of σ_i) and evaluate

$$f(\sigma_i) := \left|1 + \frac{e^{-hr_\ell}}{r_\ell}\right|$$

If $f(\sigma_i) = 0$, then r_ℓ is a root of (26.26). For 10 different values of $h \in (1.2, 2.0)$, the function $f(\sigma)$ is plotted in Fig. 26.19. This figure shows the feasible values of σ_i for which r_ℓ (defined from σ_i) is a root of (26.26).

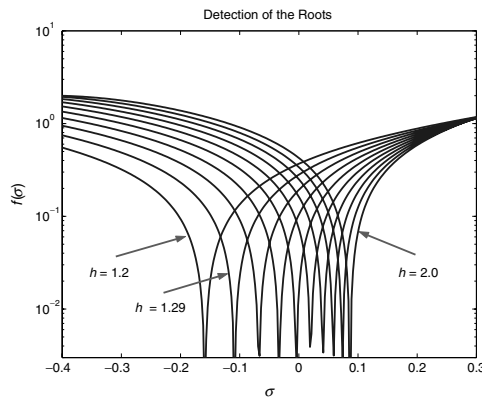



FIGURE 26.19 Detection of the dominant roots.

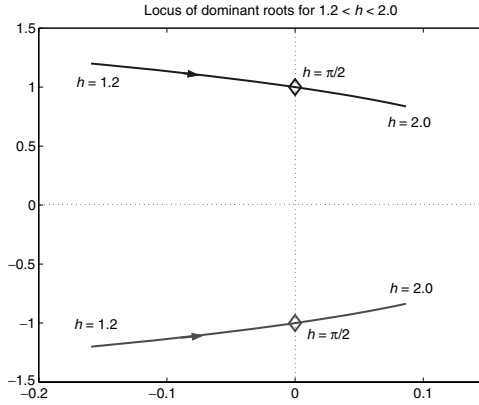



FIGURE 26.20 Dominant roots as h varies from 1.2 to 2.0.

The dominant roots of (26.26), as h varies from 1.2 to 2.0, are shown in Fig. 26.20. For $h < 1.57$, all the roots are in $\mathbb{C}_-$. For $h > 1.57$, the dominant roots are in $\mathbb{C}_+$, and for $h = 1.57$, they are at $\pm j1$.

Root Locus Using Padé Approximations

In this section we assume that $h > 0$ is fixed and we try to obtain the root locus, with respect to uncertain/adjustable gain K, corresponding to the dominant poles. The problem can be solved by numerically calculating the dominant roots of the quasi-polynomial

$$\chi(s) = D(s) + KN(s)e^{-hs} \tag{26.27}$$

for varying K, by using the methods presented in the previous section. In this section an alternative method is given that uses Padé approximation of the time delay term e^{-hs}. More precisely, the idea is to find polynomials $N_h(s)$ and $D_h(s)$ satisfying

$$e^{-hs} \approx \frac{N_h(s)}{D_h(s)} \tag{26.28}$$

so that the dominant roots

$$D(s)D_h(s) + KN(s)N_h(s) = 0 \tag{26.29}$$

closely match the dominant roots of $\chi(s)$, (26.27). How should we do the approximation (26.28) for this match?

By using the stability robustness measures determined from the Nyquist stability criterion, we can show that for our purpose we may consider the following cost function in order to define a meaningful measure for the approximation error:

$$\Delta_h =: \sup_{\omega} \left|\frac{K_{\max}N(j\omega)}{D(j\omega)}\right|\left|e^{-jh\omega} - \frac{N_h(j\omega)}{D_h(j\omega)}\right|$$

where $K_{\max}$ is the maximum value of interest for the uncertain/adjustable parameter K.

The ℓth order Padé approximation is defined as follows:

$$N_h(s) = \sum_{k=0}^{\ell} (-1)^k c_k h^k s^k$$

$$D_h(s) = \sum_{k=0}^{\ell} c_k h^k s^k$$

where coefficients c_k's are computed from

$$c_k = \frac{(2\ell - k)!\ell!}{2\ell!k!(\ell - k)!}, \quad k = 0, 1, \ldots, \ell$$

First- and second-order approximations are in the form

$$\frac{N_h(s)}{D_h(s)} = \begin{cases} \dfrac{1 - hs/2}{1 + hs/2}, & \ell = 1 \\ \dfrac{1 - hs/2 + (hs)^2/12}{1 + hs/2 + (hs)^2/12}, & \ell = 2 \end{cases}$$

Given the problem data $\{h, K_{\max}, N(s), D(s)\}$, how do we find the smallest degree, ℓ, of the Padé approximation, so that $\Delta_h \le \delta$ (or $\Delta_h / K_{\max} \le \delta'$) for a specified error δ, or a specified relative error δ'? The answer lies in the following result [7]: for a given degree of approximation ℓ we have

$$\left| e^{-jh\omega} - \frac{N_h(j\omega)}{D_h(j\omega)} \right| \le \begin{cases} 2\left(\dfrac{eh\omega}{4\ell}\right)^{2\ell+1}, & \omega \le \dfrac{4\ell}{eh} \\ 2, & \omega \ge \dfrac{4\ell}{eh} \end{cases}$$

In light of this result, we can solve the approximation order selection problem by using the following procedure:

1. Determine the frequency ω_x such that

$$\left| \frac{K_{\max} N(j\omega)}{D(j\omega)} \right| \le \frac{\delta}{2}, \quad \text{for all } \omega \ge \omega_x$$

   and initialize $\ell = 1$.
2. For each $\ell \ge 1$ define

$$\omega_\ell = \max\left\{ \omega_x, \frac{4\ell}{eh} \right\}$$

   and plot the function

$$\Phi_\ell(\omega) := \begin{cases} 2\left| \dfrac{K_{\max} N(j\omega)}{D(j\omega)} \right| \left(\dfrac{eh\omega}{4\ell}\right)^{2\ell+1}, & \text{for } \omega \le \dfrac{4\ell}{eh} \\ 2\left| \dfrac{K_{\max} N(j\omega)}{D(j\omega)} \right|, & \text{for } \omega_\ell \ge \omega \ge \dfrac{4\ell}{eh} \end{cases}$$

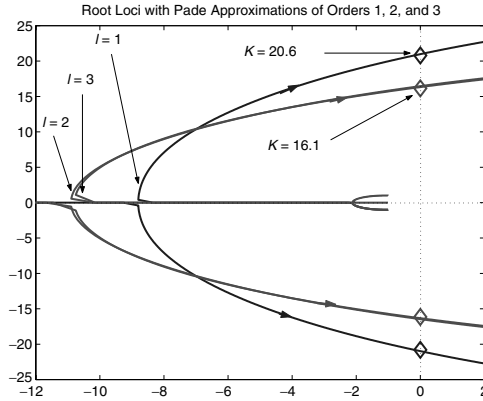



FIGURE 26.21 Dominant root for $\ell = 1$.

3. Check If

$$\max_{\omega \in [0, \omega_x]} \Phi_\ell(\omega) \le \delta \tag{26.30}$$

   If yes, stop, this value of ℓ satisfies the desired error bound: $\Delta_h \le \delta$. Otherwise, increase ℓ by 1, and go to Step 2. Note that the left-hand side of the inequality (26.30) is an upper bound of Δ_h.

Since we assumed $deg(D) > deg(N)$, the algorithm will pass Step 3 eventually for some finite $\ell \ge 1$. At each iteration, we have to draw the error function $\Phi_\ell(\omega)$ and check whether its peak value is less than δ. Typically, as δ decreases, ω_x increases, and that forces ℓ to increase. On the other hand, for very large values of ℓ, the relative magnitude c_0/c_ℓ of the coefficients becomes very large, in which case numerical difficulties arise in analysis and simulations. Also, as time delay h increases, ℓ should be increased to keep the level of the approximation error δ fixed. This is a fundamental difficulty associated with time delay systems.

Example

Let $N(s) = s + 1$, $D(s) = s^2 + 2s + 2$ and $h = 0.1$, and $K_{\max} = 20$. Then, for $\delta' = 0.05$, applying the above procedure we calculate $\ell = 2$ as the smallest approximation degree satisfying $\Delta_h/K_{\max} < \delta'$. Therefore, a second-order approximation of the time delay should be sufficient for predicting the dominant poles for $K \in [0, 20]$. Figure 26.21 shows the approximate root loci obtained from Padé approximations of degrees $\ell = 1, 2, 3$. There is a significant difference between the root loci for $\ell = 1$ and $\ell = 2$. In the region $\text{Re}(s) \ge -12$, the predicted dominant roots are approximately the same for $\ell = 2, 3$, for $K \in [0, 20]$. So, we can safely say that using higher order approximations will not make any significant difference as far as predicting the behavior of the dominant poles for the given range of K.

26.6 Notes and References

This chapter in the handbook is an edited version of related parts of the author's book [9]. More detailed discussions of the root locus method can be found in all the classical control books, such as [2, 5, 6, 8]. As mentioned earlier, extension of this method to discrete time systems is rather trivial: the method to find the roots of a polynomial as a function of a varying real parameter is independent of the variable s (in the continuous time case) or z (in the discrete time case). The only difference between these two cases is the definition of the desired region of the complex plane: for the continuous time systems, this is defined relative to the imaginary axis, whereas for the discrete time systems the region is defined with respect to the unit circle, as illustrated in Fig. 26.6.

References

1. Bellman, R. E., and Cooke, K. L., *Differential Difference Equations,* Academic Press, New York, 1963.
2. Dorf, R. C., and Bishop, R. H., *Modern Control Systems,* 9th ed., Prentice-Hall, Upper Saddle River, NJ, 2001.
3. Evans, W. R., "Graphical analysis of control systems," *Transac. Amer. Inst. Electrical Engineers,* vol. 67 (1948), pp. 547–551.
4. Evans, W. R., "Control system synthesis by root locus method," *Transac. Amer. Inst. Electrical Engineers,* vol. 69 (1950), pp. 66–69.
5. Franklin, G. F., Powell, J. D., and Emami-Naeini, A., *Feedback Control of Dynamic Systems,* 3rd ed., Addison Wesley, Reading, MA, 1994.
6. Kuo, B. C., *Automatic Control Systems,* 7th ed., Prentice-Hall, Upper Saddle River, NJ, 1995.
7. Lam, J., "Convergence of a class of Padé approximations for delay systems," *Int. J. Control,* vol. 52 (1990), pp. 989–1008.
8. Ogata, K., *Modern Control Engineering,* 3rd ed., Prentice-Hall, Upper Saddle River, NJ, 1997.
9. Özbay, H., *Introduction to Feedback Control Theory,* CRC Press LLC, Boca Raton, FL, 2000.
10. Stepan, G., *Retarded Dynamical Systems: Stability and Characteristic Functions,* Longman Scientific & Technical, New York, 1989.
11. Ulus, C., "Numerical computation of inner-outer factors for a class of retarded delay systems," *Int. J. Systems Sci.,* vol. 28 (1997), pp. 897–904.

27

Frequency Response Methods

Jyh-Jong Sheen
National Taiwan Ocean University

27.1 Introduction

The analysis and design of industrial control systems are often accomplished utilizing frequency response methods. By the term frequency response, we mean the steady-state response of a linear constant coefficient system to a sinusoidal input test signal. We will see that the response of the system to a sinusoidal input signal is also a sinusoidal output signal at the same frequency as the input. However, the magnitude and phase of the output signal differ from those of the input signal, and the amount of difference is a function of the input frequency. Thus, we will be investigating the relationship between the transfer function and the frequency response of linear stable systems.

Consider a stable linear constant coefficient system shown in Fig. 27.1. Using Euler's formula, $e^{j\omega t} = \cos\omega t + j\sin\omega t$, let us assume that the input sinusoidal signal is given by

$$u(t) = U_0 e^{j\omega t} = U_0 \cos\omega t + jU_0 \sin\omega t \tag{27.1}$$

Taking the Laplace transform of $u(t)$ gives

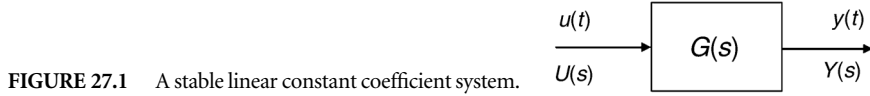
$$U(s) = \frac{U_0}{s - j\omega} = U_0 \frac{s + j\omega}{s^2 + \omega^2} = \frac{U_0 s}{s^2 + \omega^2} + j\frac{U_0 \omega}{s^2 + \omega^2} \tag{27.2}$$

The first term in Eq. (27.2) is the Laplace transform of $U_0 \cos\omega t$, while the second term, without the imaginary number j, is the Laplace transform of $U_0 \sin\omega t$.

Suppose that the transfer function $G(s)$ can be written as

$$G(s) = \frac{n(s)}{d(s)} = \frac{n(s)}{(s + p_1)(s + p_2)\cdots(s + p_n)} \tag{27.3}$$

u(t) → G(s) → y(t)

FIGURE 27.1 A stable linear constant coefficient system.

where p_i, $i = 1, 2, \ldots, n$, are assumed to be distinct poles. The Laplace transform of the output $Y(s)$ is then

$$Y(s) = G(s)U(s) = G(s)\frac{U_0}{s - j\omega} \tag{27.4}$$

Taking the partial fraction expansion of $Y(s)$ gives

$$Y(s) = \frac{k_1}{s + p_1} + \cdots + \frac{k_n}{s + p_n} + \frac{\alpha}{s - j\omega} \tag{27.5}$$

The coefficient α can be determined by

$$\alpha = [(s - j\omega)Y(s)]|_{s=j\omega} = [U_0 G(s)]|_{s=j\omega} = U_0 G(j\omega)$$

Therefore, the inverse Laplace transform of $Y(s)$ yields

$$y(t) = k_1 e^{-p_1 t} + \cdots + k_n e^{-p_n t} + U_0 G(j\omega)e^{j\omega t}, \quad t \geq 0 \tag{27.6}$$

For a stable system, all $-p_i$ have negative nonzero real parts and, therefore, all the terms $k_i e^{-p_i t}$, $i = 1, 2, \ldots, n$, approach zero as t approaches infinity. Thus, at steady state, the output $y(t)$ becomes

$$y_{ss}(t) = \lim_{t \to \infty} y(t) = U_0 G(j\omega)e^{j\omega t} = U_0|G(j\omega)|e^{j(\omega t + \phi)} \tag{27.7}$$

The sinusoidal transfer function, $G(j\omega)$, is written in exponential form

$$G(j\omega) = |G(j\omega)|e^{j\phi}$$

where

$$|G(j\omega)| = \sqrt{\{\text{Re}[G(j\omega)]\}^2 + \{\text{Im}[G(j\omega)]\}^2} \tag{27.8a}$$

and

$$\phi = \angle G(j\omega) = \tan^{-1}\frac{\text{Im}[G(j\omega)]}{\text{Re}[G(j\omega)]} \tag{27.8b}$$

Equation (27.7) shows that for a stable system subject to a sinusoidal input, the steady-state response is a sinusoidal output of the same frequency as the input. The amplitude of the output is that of the input times $|G(j\omega)|$, and the phase angle differs from that of the input by the amount $\phi = \angle G(j\omega)$.

Example 1

A first-order low-pass filter is shown in Fig. 27.2. The transfer function of this filter is

$$G(s) = \frac{V_o(s)}{V_i(s)} = \frac{1}{RCs + 1}$$

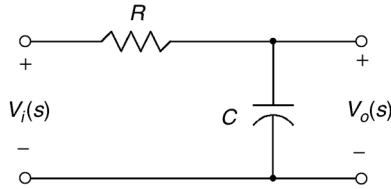



FIGURE 27.2 A first-order low-pass filter.

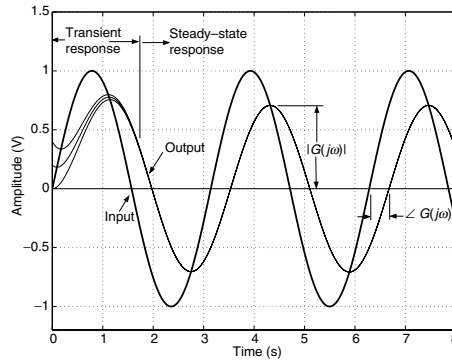



FIGURE 27.3 Frequency response of $G(s) = 1/(0.5s + 1)$ to $u(t) = \sin 2t$.

The sinusoidal transfer function is given by

$$G(j\omega) = \frac{1}{j\omega(RC)+1} = \frac{1}{j(\omega/\omega_1)+1}$$

where $\omega_1 = 1/RC$. The magnitude and phase angle of the frequency response are

$$|G(j\omega)| = \frac{1}{\sqrt{1+(\omega/\omega_1)^2}} \qquad \text{and} \qquad \phi(\omega) = -\tan^{-1}\frac{\omega}{\omega_1}$$

Figure 27.3 shows the response of the system with $RC = 0.5$ to the input $u = \sin 2t$. It can be seen that the steady-state response is irrelevant to the initial conditions, and the steady-state amplitude of the output is $1/\sqrt{2}$ and the phase angle is $-45°$.

27.2 Bode Plots

There are three commonly used displays of frequency response of a system. They are:

1. the Bode diagram or logarithmic plot,
2. the polar plot, and
3. the log-magnitude versus phase plot or Nichols chart.

We will present Bode diagrams of sinusoidal transfer functions in this section, followed by the sections on polar plots and log-magnitude versus phase plots.

The main advantages in using the logarithmic plot are the capability of plotting low and high frequency characteristics of the transfer function in one diagram, and the relative ease of adding the separate terms

of a high-order transfer function graphically. The basic types of factors that may occur in a transfer function are as follows:

1. constant gain K,
2. poles (or zeros) at the origin $(j\omega)^{\pm n}$,
3. poles (or zeros) on the real axis $(j\omega\tau + 1)^{\pm 1}$, and
4. complex conjugate poles (or zeros) $[(j\omega/\omega_n)^2 + 2\varsigma(j\omega/\omega_n) + 1]^{\pm 1}$.

The curves of logarithmic magnitude and phase angle for these four factors can easily be drawn and then added together graphically to obtain the curves for the complete transfer function. The process of drawing the logarithmic plot can be further simplified by using asymptotic approximations to these curves and obtaining the actual curves at specific important frequencies.

Constant Gain K

The logarithmic gain for the constant gain K is

$$20\log|K| = \text{constant in decibel}, \quad \angle K = \begin{cases} 0°, & \text{if } K > 0 \\ -180°, & \text{if } K < 0 \end{cases}$$

The gain and phase curves are simply horizontal lines on the Bode diagram.

Poles (or Zeros) at the Origin $(j\omega)^{\pm n}$

Since

$$20\ \log|j\omega|^{\pm n} = \pm 20 n \log\omega, \quad \angle(j\omega)^{\pm n} = \pm n \times 90°$$

the slopes of the magnitude curves are $\pm 20n$ dB/decade for the factor $(j\omega)^{\pm n}$ and the phase angles are constants equal to $\pm n \times 90°$.

Poles (or Zeros) on the Real Axis $(j\omega\tau + 1)^{\pm 1}$

For a pole factor $(j\omega\tau + 1)^{-1}$,

$$\left|\frac{1}{j\omega\tau + 1}\right| = \frac{1}{\sqrt{\omega^2\tau^2 + 1}}$$

The magnitude of the pole factor is 1 when $\omega \ll 1/\tau$, and $1/(\omega\tau)$ when $\omega \gg 1/\tau$. Thus, there are two asymptotic curves for the pole factor,

$$20\log\left|\frac{1}{j\omega\tau + 1}\right| \approx \begin{cases} 0\ \text{dB}, & \text{when } \omega \ll \dfrac{1}{\tau} \\ -20\log\omega\tau = -20\left(\log\omega - \log\dfrac{1}{\tau}\right), & \text{when } \omega \gg \dfrac{1}{\tau} \end{cases}$$

The slope of the asymptotic curve when $\omega \gg 1/\tau$ is -20 dB/decade for the pole factor. The two asymptotes intersect at $\omega = 1/\tau$, the break frequency or the corner frequency. The actual logarithmic gain at $\omega = 1/\tau$ is -3 dB. The phase angle is $\phi(\omega) = -\tan^{-1}\omega\tau$.

The Bode diagram of a zero factor $(j\omega\tau + 1)$ is obtained in the same manner. However, the slope of the magnitude asymptotic curve when $\omega \gg 1/\tau$ is +20 dB/decade, and the phase angle is $\phi(\omega) = +\tan^{-1}\omega\tau$. The Bode diagrams of first-order factors are shown in Fig. 27.4. Linear approximations to the phase angle curves are also presented.

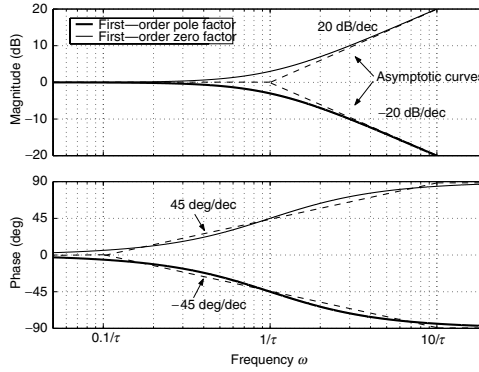



FIGURE 27.4 Bode diagrams for the first-order factors $(j\omega\tau + 1)^{\pm 1}$.

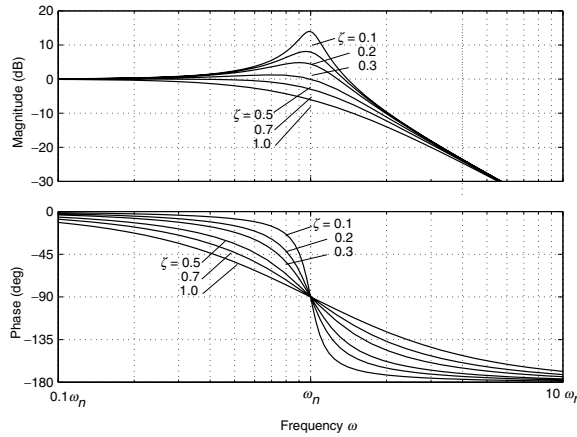



FIGURE 27.5 Bode diagram for the quadratic pole factor $[(j\omega/\omega_n)^2 + 2\zeta(j\omega/\omega_n) + 1)]^{-1}$.

Complex Conjugate Poles (or Zeros) $[(j\omega/\omega_n)^2 + 2\varsigma(j\omega/\omega_n) + 1]^{\pm 1}$

The magnitude and phase angle of the complex conjugate poles $[(j\omega/\omega_n)^2 + 2\zeta(j\omega/\omega_n) + 1]^{-1}$ are

$$\left|\left(j\frac{\omega}{\omega_n}\right)^2 + 2\varsigma\left(j\frac{\omega}{\omega_n}\right) + 1\right|^{-1} = \left[\left(1 - \frac{\omega^2}{\omega_n^2}\right)^2 + \left(2\varsigma\frac{\omega}{\omega_n}\right)^2\right]^{-1/2}$$

$$\angle\left[\left(j\frac{\omega}{\omega_n}\right)^2 + 2\varsigma\left(j\frac{\omega}{\omega_n}\right) + 1\right]^{-1} = -\tan^{-1}\frac{2\varsigma\omega/\omega_n}{1 - \omega^2/\omega_n^2}$$

The magnitude of the complex conjugate pole factor is 1 when $\omega \ll \omega_n$, and $(\omega/\omega_n)^{-2}$ when $\omega \gg \omega_n$. Therefore, the two asymptotic curves for the complex conjugate pole factor are

$$20\log\left|\left(j\frac{\omega}{\omega_n}\right)^2 + 2\varsigma\left(j\frac{\omega}{\omega_n}\right) + 1\right|^{-1} \approx \begin{cases} 0 \text{ dB}, & \text{when } \omega \ll \omega_n \\ -40(\log\omega - \log\omega_n), & \text{when } \omega \gg \omega_n \end{cases}$$

The slope of the asymptotic curve when $\omega \gg \omega_n$ is –40 dB/decade for the complex conjugate pole factor. The magnitude asymptotes intersect at $\omega = \omega_n$, the natural frequency. The actual gain at $\omega = \omega_n$ is $G(j\omega_n) = 1/2\zeta$. The Bode diagram of a complex conjugate pole factor is shown in Fig. 27.5. It is seen from Fig. 27.5 that the

difference between the actual magnitude curve and the asymptotic approximation is a function of damping ratio. The resonant frequency ω_r is defined as the frequency where the peak value of the frequency response M_r occurs. When the damping ratio approaches zero, ω_r approaches ω_n. The resonant frequency can be determined by taking the derivative of the magnitude with respect to the frequency, and setting it equal to zero. The resonant frequency and the peak value of the magnitude are represented by

$$\omega_r = \omega_n\sqrt{1-2\varsigma^2}, \quad \varsigma < 0.707 \tag{27.9a}$$

and

$$M_r = \frac{1}{2\varsigma\sqrt{1-\varsigma^2}}, \quad \varsigma < 0.707 \tag{27.9b}$$

Example 2

Let us consider the transfer function

$$G(s) = \frac{10(s/5+1)}{s(s+1)[(s/10)^2+(s/10)+1]}$$

We first list the basic factors of $G(s)$ in Table 27.1 in the order of increasing corner or natural frequencies.

The complete asymptotic magnitude curve for $G(j\omega)$ is produced by adding together the asymptotic logarithmic magnitudes of each factor, as shown by the solid line in Fig. 27.6. Since the dc gain of each factor is 1, these factors have no effect on the asymptotic magnitude until the frequency approaches their corner or natural frequencies. Thus, the asymptotic magnitude can be quickly obtained by plotting each asymptote in order as frequency increases. The asymptotic curve intersects 20 dB at $\omega = 1$ with the slope –20 dB/decade due to the pole at the origin and the constant gain $K = 10$. At $\omega = 1$ the slope further decreases to –40 dB/decade due to the pole at $\omega = 1$. Then at $\omega = 5$ the slope increases to –20 dB/decade

TABLE 27.1 The Basic Factors of $G(j\omega)$

Type of Factors	Constant Gain	Pole	Pole	Zero	Complex Poles
Corner frequency	$K = 10$	0	1	5	10
Order	0	–1	–1	+1	–2

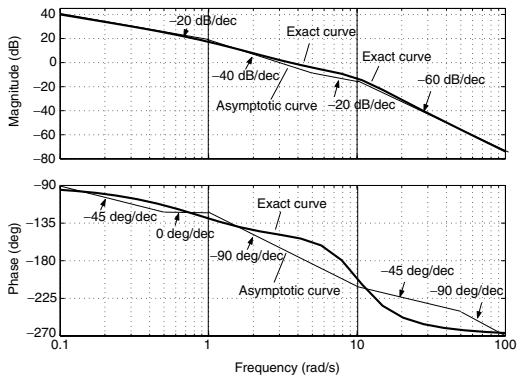



FIGURE 27.6 The Bode plot of the transfer function in Exmple 2.

due to the zero at $\omega = 5$. Finally at $\omega = 10$ the slope becomes –60 dB/decade due to the complex conjugate poles at $\omega_n = 10$.

The exact magnitude is obtained by calculating the actual magnitude at important frequencies such as the corner or natural frequencies of each factor. The phase curve can be obtained by adding the phase due to each factor. Although the linear approximation of the phase characteristic for a single pole or zero is suitable for initial analysis, the error between the exact phase curve and the linear approximation of complex conjugate poles can be large, as seen in Fig. 27.6. Hence, if the accurate phase angle curve is required, a computer program such as Matlab or Ctrl-C can be utilized to generate the actual phase curve.

27.3 Polar Plots

The polar plot of a sinusoidal transfer function $G(j\omega)$ is a plot of both the magnitude and the phase of the frequency response in polar coordinates as the frequency ω varies from zero to infinity. Since the sinusoidal transfer function $G(j\omega)$ can be expressed as

$$G(j\omega) = \mathrm{Re}[G(j\omega)] + j\,\mathrm{Im}[G(j\omega)] = |G(j\omega)|e^{j\phi}$$

the polar plot of $G(j\omega)$ is a plot of $\mathrm{Re}[G(j\omega)]$ on the horizontal axis versus $\mathrm{Im}[G(j\omega)]$ on the vertical axis in the complex $G(s)$-plane as ω varies from zero to infinity. Hence, for each value of ω, a polar plot of $G(j\omega)$ is defined by a vector of length $|G(j\omega)|$ and a phase angle $\phi = \angle G(j\omega)$, as in Eq. (27.8).

We can investigate the general shapes of polar plots according to the system types and relative degrees of transfer functions. Relative degree of a transfer function is defined as the difference between the degree of the denominator polynomial and that of the numerator. Consider a transfer function of the form

$$\begin{aligned} G(j\omega) &= \frac{K(1 + j\omega\tau_a)(1 + j\omega\tau_b)\cdots}{(j\omega)^N(1 + j\omega\tau_1)(1 + j\omega\tau_2)\cdots} \\ &= \frac{b_0(j\omega)^m + b_1(j\omega)^{m-1} + \cdots}{a_0(j\omega)^n + b_1(j\omega)^{n-1} + \cdots} \end{aligned}$$

where $K > 0$ and the relative degree $n - m \geq 0$. The magnitudes and phase angles of $G(j\omega)$ as ω approaches zero and infinity are presented in Table 27.2. The general shapes of the polar plots of various system types in the low-frequency portion are shown in Fig. 27.7. The high-frequency portions of the polar plots of various relative degrees are shown in Fig. 27.8. It can be seen that the $G(j\omega)$ loci are parallel to either the horizontal or the vertical axes with infinite magnitude as $\omega \to 0^+$ for system types greater than zero. If the relative degree is greater than zero, the $G(j\omega)$ loci converge to the origin clockwise and are tangent to one or the other axes. Note that the polar plot curves can be very complicated due to the numerator and denominator dynamics over the intermediate frequency range. Therefore, the polar plot of $G(j\omega)$ in the frequency range of interest must be accurately determined.

TABLE 27.2 $G(j\omega)$ vs. System Type and Relative Degree as $\omega \to 0^+$ and ∞

System Type N	$\omega \to 0^+$	Relative Degree $n - m$	$\omega \to \infty$
0	$K\angle 0°$	0	$b_0/a_0 \angle 0°$
1	$\infty\angle -90°$	1	$0\angle -90°$
2	$\infty\angle -180°$	2	$0\angle -180°$
3	$\infty\angle -270°$	3	$0\angle -270°$

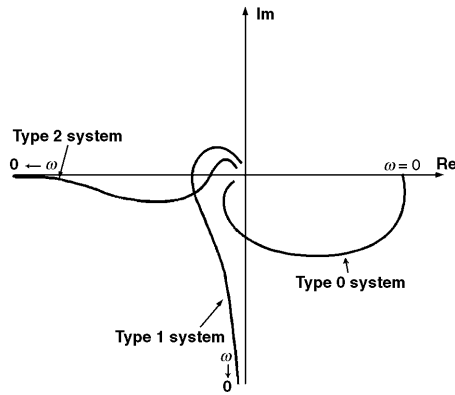



FIGURE 27.7 Polar plots of system with various system types as $\omega \rightarrow 0$.

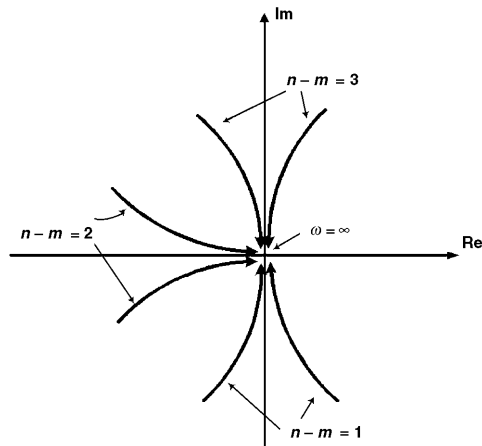



FIGURE 27.8 Polar plots of system with various relative degrees as $\omega \rightarrow \infty$.

We will see that for a closed-loop system, the polar plot of the loop transfer function is useful in determining the stability of the system. The polar plots of some simple systems are shown in Fig. 27.9.

27.4 Log-Magnitude Versus Phase Plots

Another approach to presenting the frequency response of a system by a single graph is to plot its logarithmic magnitude versus the phase angle over a frequency range of interest. The resulting curve is a function of the frequency ω. Such log-magnitude versus phase plots are called Nichols charts.

Advantages of the Nichols chart are that the relative stability of the closed-loop system can be determined quickly and that the process of closed-loop compensation can be carried out easily. The Nichols charts of the systems in Fig. 27.9 are depicted in Fig. 27.10 for comparison. Figure 27.11 displays three different frequency-response curves of the second-order system

$$G(s) = \frac{\omega_n^2}{s^2 + 2\varsigma\omega_n s + \omega_n^2}$$

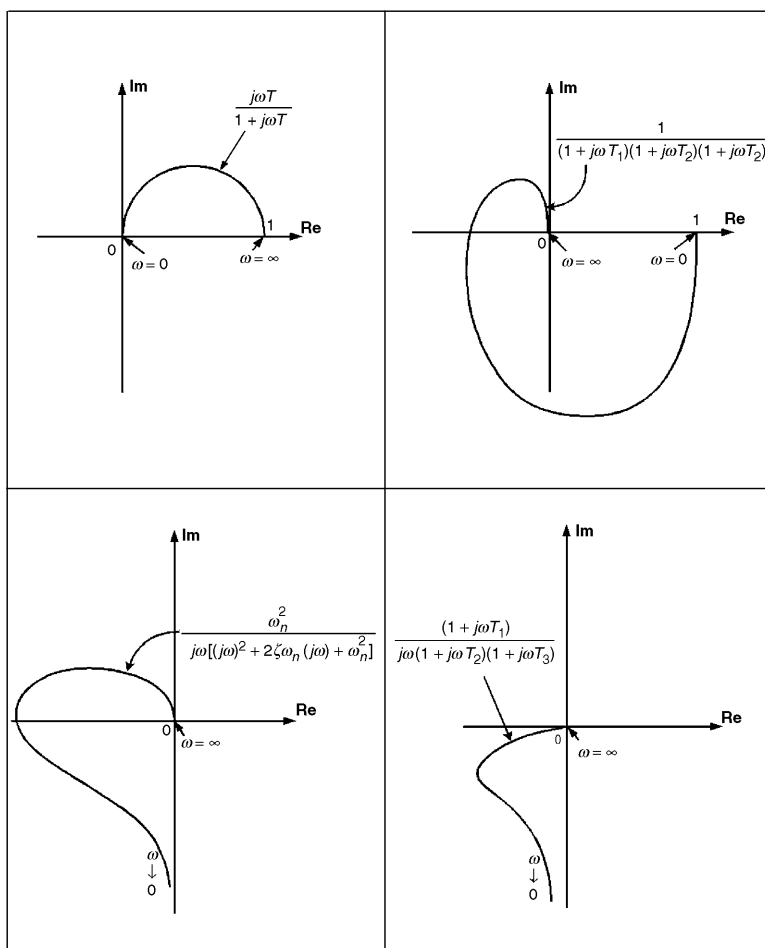



FIGURE 27.9 Polar plots of simple transfer functions.

27.5 Experimental Determination of Transfer Functions

We can obtain a transfer function model from frequency-response measurements of a stable system. First, the Bode diagram of the frequency response is plotted from the measurements. Then the open-loop transfer function can be deduced from the magnitude and phase plots based on the relationships of the basic pole and zero factors.

A wave analyzer is a device to measure the amplitudes and phases of the steady-state response as the frequency of the input sinusoidal wave is altered. A transfer function analyzer can be used to measure the open-loop and closed-loop transfer functions.

We will use a computer program combined with an analog-to-digital and digital-to-analog (AD and DA) card to generate the sinusoidal input signal and to measure the frequency response of a system. Consider the second-order Sallen-Key low-pass filter in Fig. 27.12. The transfer function of the filter is given by

$$G(s) = \frac{V_o(s)}{V_i(s)} = \frac{K}{s^2/\omega_n^2 + 2\varsigma(s/\omega_n) + 1} \tag{27.10}$$

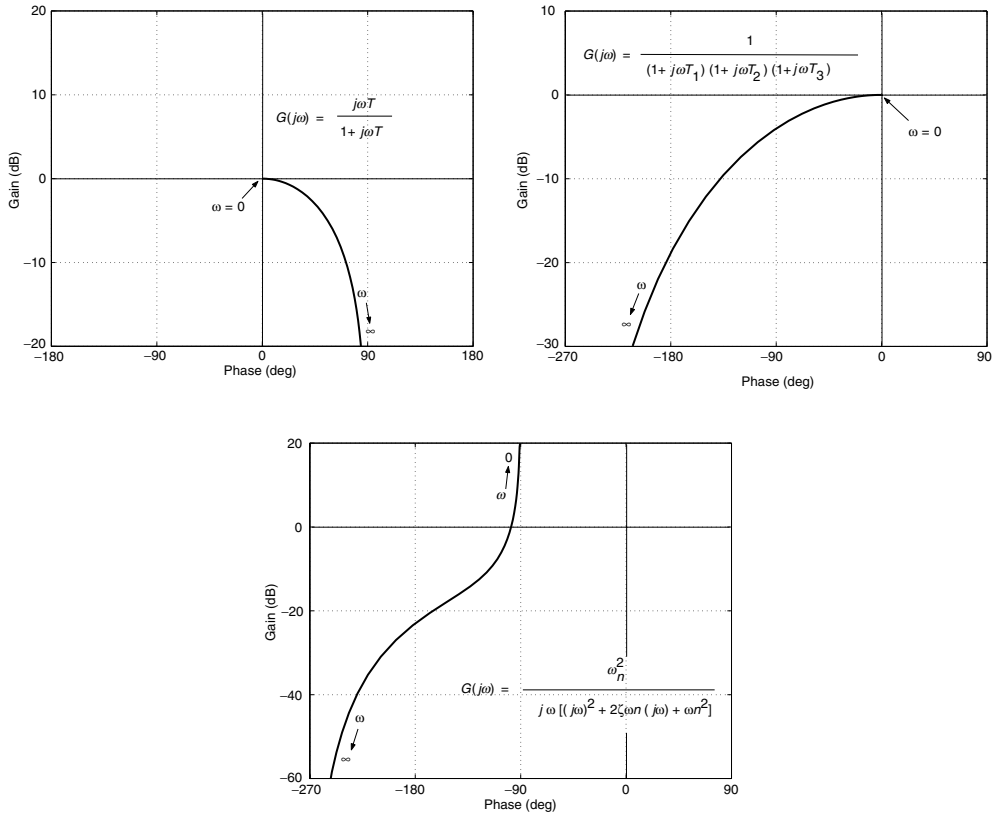



FIGURE 27.10 Nichols charts of simple transfer functions.

where

$$K = \frac{R_1 + R_2}{R_2}, \quad \omega_n = \frac{1}{\sqrt{R_A R_B C_A C_B}}$$

and

$$\varsigma = \frac{1}{2}\left[\sqrt{\frac{C_B}{C_A}}\frac{R_A + R_B}{\sqrt{R_A R_B}} + (1-K)\sqrt{\frac{R_A C_A}{R_B C_B}}\right]$$

The Real-Time Windows Target in Matlab is used with an Advantech PCL-818L AD and DA card. The sampling time is 0.001 s. The measured magnitudes and phase angles are shown in Fig. 27.13. From the Bode plot, we can find that the dc gain is equal to 1.995 and the natural frequency ω_n = 17.90 rad/s. From Eq. (27.9b) and M_r = 1.993, we have ζ = 0.26.

An alternative to estimating the transfer function is to use an excitation signal that is sufficiently rich in the frequency contents of interest and to measure the corresponding output. System identification technique is then applied to find the order and parameters of the transfer function. Suitable excitation signals are the impulse signal, sweep sine signal, random sequence, and so forth. Figure 27.14 presents the sweep sine input and the corresponding output. The Matlab System Identification Toolbox is then

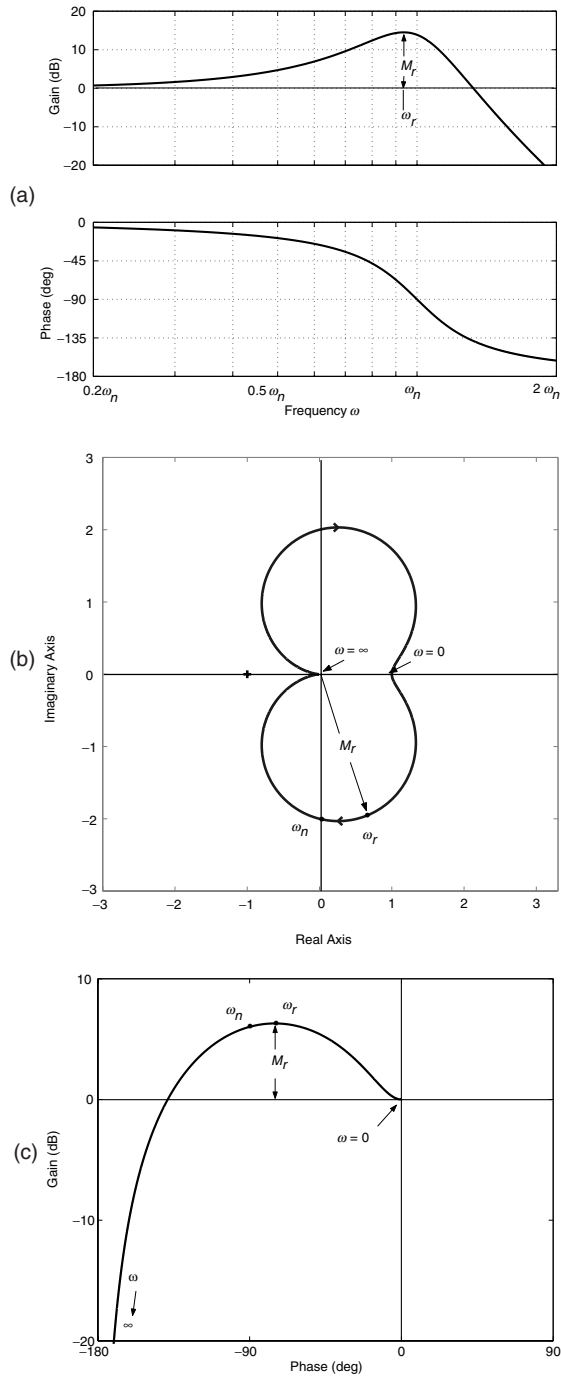



FIGURE 27.11 Three frequency response representations of $G(s) = \omega_n^2/(s^2 + 2\zeta\omega_n s + \omega_n^2)$: (a) Bode diagram, (b) polar plot, and (c) Nichols chart.

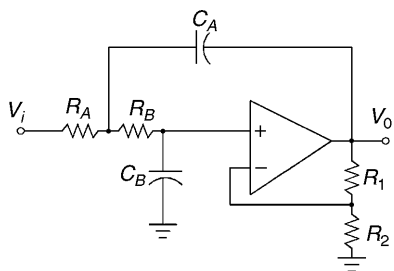



FIGURE 27.12 Sallen-Key low-pass filter.

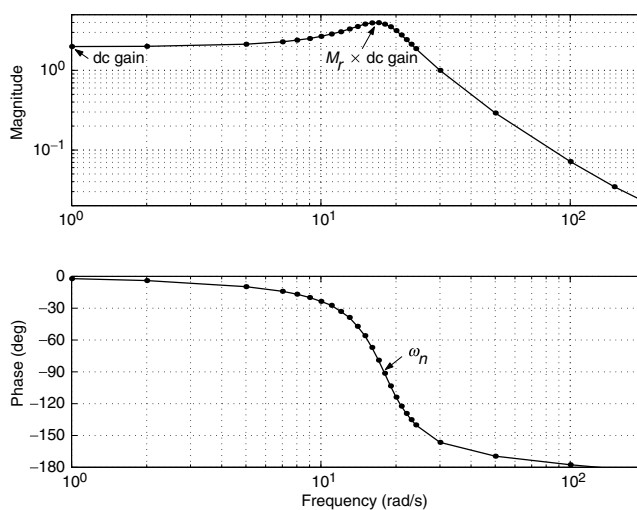



FIGURE 27.13 Frequency response of the Sallen-Key filter from experimental data.

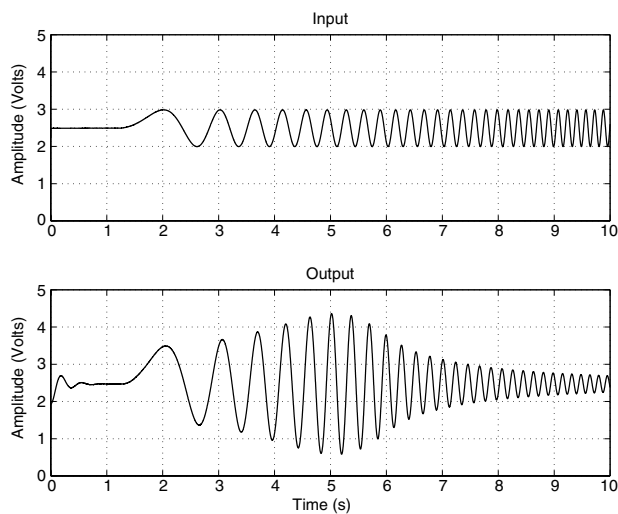



FIGURE 27.14 Sweep sine response of the Sallen-Key filter.

TABLE 27.3 Estimated Transfer Functions for the Second-Order Low-Pass Filter

1. Ideal op-amp circuit: $G(s) = \dfrac{1.997}{(s/18.09)^2 + 2\times 0.271(s/18.09) + 1}$ where the measured values of resistors and capacitors are substituted in Eq. (27.10) with $R_1 =$ 98.4 kΩ, $R_2 = 98.7$ kΩ, $R_A = 51.3$ kΩ, $R_B = 98.5$ kΩ, $C_A = 1.083$ μF, and $C_B = 0.564$ μF.
2. From the Bode plot: $G(s) = \dfrac{1.995}{(s/17.90)^2 + 2\times 0.259(s/17.90) + 1}$
3. System identification: $G(s) = \dfrac{1.997}{(s/17.78)^2 + 2\times 0.255(s/17.78) + 1}$

utilized to estimate the transfer function. The resulting transfer functions from the ideal op-amp circuit in Eq.(27.10), the Bode plot, and system identification are shown in Table 27.3 for comparison. It is seen that the differences among the three transfer functions are very small. However, the task of determining transfer functions from Bode plots can be very difficult as various pole or zero factors of close corner frequencies can complicate the magnitude and phase plots for high-order systems. Thus, it is recommended that system identification technique be used for determination of high-order transfer functions.

27.6 The Nyquist Stability Criterion

The Nyquist stability criterion provides a graphical procedure for determining the closed-loop stability from the open-loop frequency-response curves. The criterion is based on a result from complex variables theory known as the argument principle, due to Cauchy.

Suppose $F(s)$ is a rational function of s with real coefficients that are analytic everywhere in the s-plane except at its poles. Let Γ_s be a closed, clockwise contour in the s-plane that does not pass through any zeros or poles of $F(s)$. The contour map Γ_F is defined by substituting the values of s on the contour Γ_s for s in $F(s)$. The resulting map is also a closed continuous contour in the $F(s)$-plane. The principle of the argument can be stated as follows:

> A contour map Γ_F of a complex function $F(s)$ defined on Γ_s in the s-plane will only encircle the origin of the $F(s)$-plane if the contour contains a pole or zero of the function. The net number that Γ_F encircles the origin in the clockwise direction is
>
> $$N = Z - P \tag{27.11}$$
>
> where Z and P are, respectively, the numbers of zeros and poles of $F(s)$ enclosed by a closed clockwise contour Γ_s in the s-plane.

Example 3

To illustrate the argument principle, consider a rational function

$$F(s) = \frac{(s+3)(s+4)}{(s+1)(s+2)}$$

which has zeros at $s = -3, -4$ and poles at $s = -1, -2$. The various contour maps of $F(s)$ are shown in Fig. 27.15, where Γ_r denotes the contour map of a clockwise circular contour of radius r in the s-plane. We have the following observations from Fig. 27.15:

1. The contour map $\Gamma_{0.5}$ does not encircle the origin of the $F(s)$-plane as the contour in the s-plane does not encircle any pole or zero.
2. $\Gamma_{1.99}$ encircles the origin once in the counterclockwise direction as the contour encircles the pole at $s = -1$ in the clockwise direction in the s-plane, and from Eq. (27.11), $N = Z - P = 0 - 1 = -1$. Note that $\Gamma_{1.99}$ is a closed contour with two loops and only the one encircling the origin is shown in Fig. 27.15.

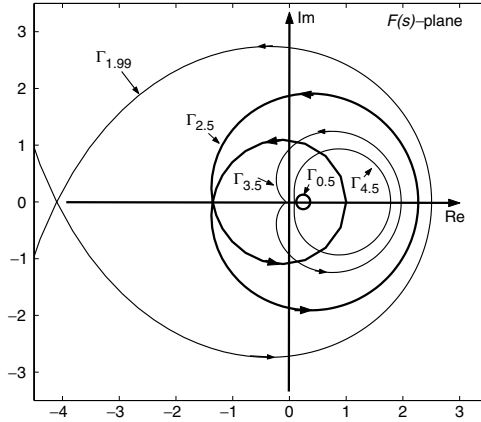



FIGURE 27.15 The contour maps of $F(s)$ in Example 3.

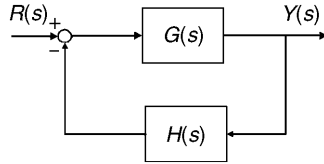



FIGURE 27.16 Closed-loop system.

3. $\Gamma_{2.5}$ encircles the origin twice in the counterclockwise direction as the contour contains two poles at $s = -1, -2$ and $N = Z - P = 0 - 2 = -2$.
4. When the radius of the contour is increased to contain the poles at $s = -1, -2$ and the zero at $s = -3$, then $N = Z - P = 1 - 2 = -1$ and a contour map like $\Gamma_{3.5}$ encircles the origin once in the counterclockwise direction.
5. When the radius of the contour is further increased to encircle the two poles and two zeros, then $N = 2 - 2 = 0$ and the contour map like $\Gamma_{4.5}$ does not encircle the origin.

We now apply Cauchy's principle of argument to develop the Nyquist stability criterion. Suppose that the characteristic equation of the closed-loop system in Fig. 27.16 is

$$F(s) = 1 + G(s)H(s) = 0$$

Let $L(s) = G(s)H(s)$, the loop transfer function. Using the argument principle, let us assume that none of the poles or zeros of $F(s)$ lie on the imaginary axis in the s-plane. We now define the Nyquist path, Γ_s, that is composed of the imaginary axis and a semicircle of infinite radius. This contour completely encloses the entire complex right-half plane as depicted in Fig. 20.17(a). The corresponding contour map Γ_F is shown in Fig. 27.17(b). It follows from the argument principle that N corresponds to the net number of clockwise encirclements of the origin of the $1 + L(s)$-plane by Γ_F. P is the number of poles of $F(s)$ in the right-half s-plane and thus is the number of poles of the loop transfer function $L(s)$ in the right-half s-plane. Z is the number of zeros of the characteristic equation $F(s)$ of the closed-loop system in the right-half s-plane. Therefore, Z must be zero for the closed-loop system to be stable.

In practice, a modification is made to simplify the application of the Nyquist criterion. Instead of plotting Γ_F in the $1 + L(s)$-plane, we plot just $L(s)$ evaluated along the contour Γ_s. The resulting contour map Γ_L is in the $L(s)$-plane and has the same shape as Γ_F but is shifted 1 unit to the left, as shown in Fig. 27.17(c). It thus follows that N is the net number of encirclements of the -1 point in the $L(s)$-plane.

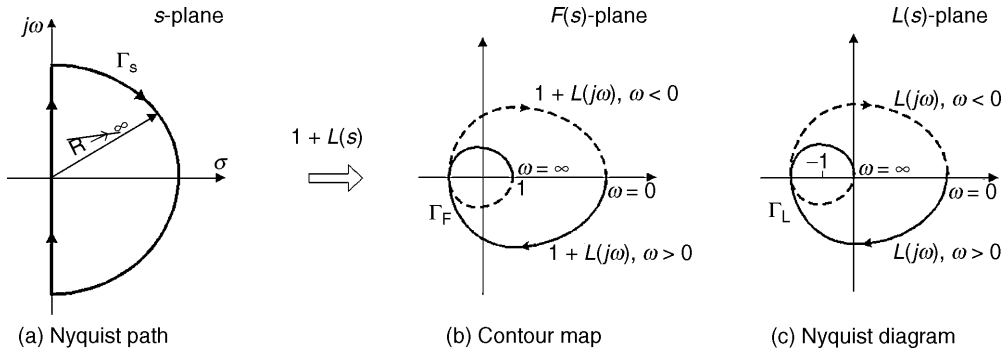



FIGURE 27.17 Nyquist diagram.

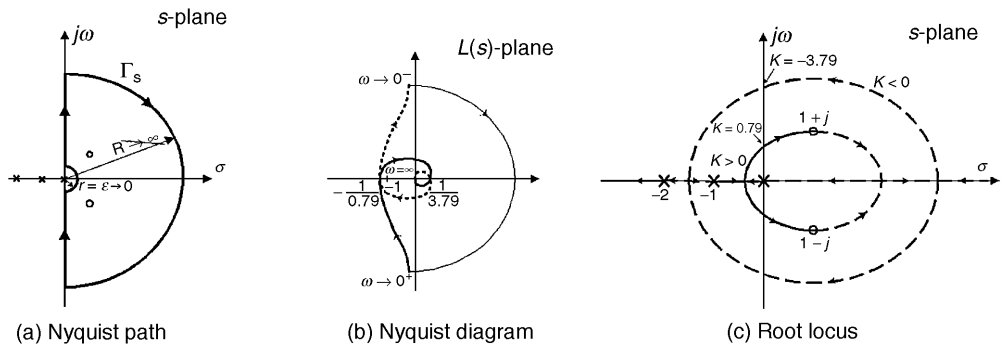



FIGURE 27.18 Nyquist diagram and root locus of Example 4.

The Nyquist stability criterion can now be stated as follows:

> A necessary and sufficient condition for the closed-loop stability of a system defined by the loop transfer function $L(s)$ is that
>
> $$Z = N + P \tag{27.12}$$
>
> be equal to zero, where N is the net number of encirclements of the -1 point in the $L(s)$-plane, and P is the number of unstable poles of the loop transfer function $L(s)$.

Example 4

Consider the system with the loop transfer function

$$KL(s) = KG(s)H(s) = K\frac{s^2 - 2s + 2}{s(s+1)(s+2)} \tag{27.13}$$

Let us determine the range of the gain K such that the closed-loop system is stable. Since there is a pole at $s = 0$, we need to modify the Nyquist path to detour around the origin. The contour is shown in Fig. 27.18(a), where the detour is chosen to be a semicircle of radius approaching zero in the limit. We use the following procedure to sketch the Nyquist plot in Fig. 27.18(b):

1. Determine $L(j\omega)$ as $\omega \to 0^+$: $L(s)$ is of system type 1 and thus

$$L(j\omega) \approx \frac{1}{j\omega} = \infty\angle{-90^\circ}$$

according to Table 27.2.

2. Determine $L(j\omega)$ as $\omega \to \infty$: $L(s)$ has a relative degree of 1 and

$$L(j\omega) \approx \frac{1}{j\omega} = 0\angle -90^\circ$$

according to Table 27.2.

3. From the Bode plot, draw the polar plot of $L(j\omega)$ as ω varies from 0^+ to ∞. Although the magnitude curve of the factor $(s^2 - 2s + 2)$ is the same as the factor $(s^2 + 2s + 2)$, the phase of the factor $(s^2 - 2s + 2)$ changes from 0° to −180°. Thus, a sketch of the Bode diagram shows that the magnitude curve varies from infinity to zero and the phase changes from −90° to −450°. Since there are two points at which the phases are −180° and −360°, there will be two intersections of the $L(j\omega)$ locus with the real axis in the $L(s)$-plane.
4. Draw the polar plot of $L(j\omega)$ as ω varies from 0^- to $-\infty$ by reflecting the curve of $L(j\omega)$ in procedure 3 with respect to the real axis in the $L(s)$-plane.
5. Determine the contour map of the small detour around the origin of the s-plane to complete the plot. On the detour,

$$s = \lim_{\varepsilon \to 0} \varepsilon e^{j\theta}, \quad -90^\circ \le \theta \le 90^\circ$$

The contour map of the detour can then be determined by

$$\lim_{\varepsilon \to 0} L(\varepsilon e^{j\theta}) = \lim_{\varepsilon \to 0} \frac{(\varepsilon e^{j\theta})^2 - 2\varepsilon e^{j\theta} + 2}{\varepsilon e^{j\theta}(\varepsilon e^{j\theta} + 1)(\varepsilon e^{j\theta} + 2)} = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon e^{j\theta}} = \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \angle -\theta$$

The resulting map is a large semicircle of radius approaching infinity. This semicircle starts at the point $L(j0^-)$ and swings 180° in the counterclockwise direction to connect the point $L(j0^+)$ in the $L(s)$-plane.

6. Calculate the intersections of the $L(j\omega)$ locus with the real-axis, for these points are related to the relative stability of the system. Suppose that the $L(j\omega)$ locus intersects the real axis for some critical frequency ω_{cr}. Then

$$L(j\omega_{cr}) = \begin{cases} 180^\circ + k360^\circ, & \text{for } K > 0 \\ 0^\circ + k360^\circ, & \text{for } K < 0 \end{cases}$$

where $k = 0, \pm 1, \pm 2, \pm 3, \ldots$. This phase condition at the critical frequency is directly related to the angle condition of the root locus when the root locus crosses the imaginary axis. Therefore, we can utilize the Routh–Hurwitz criterion to determine the points where the $L(j\omega)$ locus crosses the real axis. The characteristic equation of the system (27.13) can be written as

$$s^3 + (K+3)s^2 + (2-2K)s + 2K = 0$$

Thus, the Routh array is

$$\begin{array}{c|ll} s^3 & 1 & 2-2K \\ s^2 & K+3 & 2K \\ s^1 & c & \\ s^0 & 2K & \end{array}$$

where

$$c = \frac{(K+3)(2-2K) - 2K}{K+3}$$

Let $c = 0$, and solving for K, we get the critical gains

$$K_{cr} = \frac{-3 \pm \sqrt{21}}{2} = 0.79, -3.79$$

Substituting the values of K_{cr} in the auxiliary equation

$$(K_{cr} + 3)s^2 + 2K_{cr} = 0$$

we obtain the critical frequencies

$$\omega_{cr} = \sqrt{\frac{2K_{cr}}{K_{cr} + 3}} = \begin{cases} 0.65, & \text{when } K_{cr} = 0.79 \\ 3.10, & \text{when } K_{cr} = -3.79 \end{cases}$$

At the critical frequency, we have the characteristic equation

$$1 + K_{cr} L(j\omega_{cr}) = 0$$

Hence the points of the $L(j\omega)$ locus that cross the real-axis are

$$L(j\omega_{cr}) = -\frac{1}{K_{cr}} = -\frac{1}{0.79}, \frac{1}{3.79}$$

The complete Nyquist plot is shown not to scale in Fig. 27.18(b). The range of the gain K for which the system is stable can be determined using Nyquist criterion. For different values of K, the Nyquist diagram needs to be redrawn in order to count the number of encirclement of the -1 point. We can avoid this by counting the number of encirclement of $-1/K$ point instead. From the Nyquist criterion, $Z = N + P$, where $P = 0$. It can be seen from Fig. 27.18(b) that there are four cases of the encirclements of the $-1/K$ point.

1. $K > 0$ and $-1/K < -1/0.79 \Rightarrow 0 < K < 0.79$, and $N = 0$. We have $Z = 0$ and the system is stable.
2. $K > 0$ and $-1/K > -1/0.79 \Rightarrow K > 0.79$, and $N = 2$. We have $Z = 2$ and the system has two unstable poles.
3. $K < 0$ and $-1/K < 1/3.79 \Rightarrow K < -3.79$, and $N = 3$. We have $Z = 3$ and the system has three unstable poles.
4. $K < 0$ and $-1/K > 1/3.79 \Rightarrow -3.79 < K < 0$ and $N = 1$. We have $Z = 1$ and the system has one unstable pole.

The root locus of system (27.13) is also shown in Fig. 27.18(c) for comparison.

27.7 Relative Stability

In designing a control system, it is required that the system be stable. In addition to stability, there are important concerns such as acceptable transient response and capability to deal with model uncertainty. Since the model used in the design and analysis of a control system is never exact, it may suggest a stable system; but the physical system turns out to be unstable. Therefore, it is required that the system not only be stable but also have some stability margin or adequate relative stability.

Suppose that the sinusoidal loop transfer function $L(j\omega)$ locus intersects the -1 point for some critical frequency ω_{cr}. Then

$$L(j\omega_{cr}) = G(j\omega_{cr})H(j\omega_{cr}) = -1$$

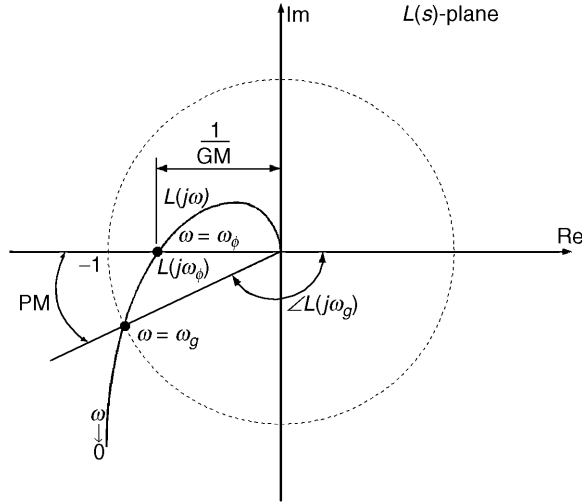



FIGURE 27.19 Gain and phase margins.

or

$$1 + L(j\omega_{\text{cr}}) = 1 + G(j\omega_{\text{cr}})H(j\omega_{\text{cr}}) = 0$$

This indicates that the closed-loop system has a pair of complex poles at $s = \pm j\omega_{cr}$. Hence, the system is marginally stable and oscillates with the frequency ω_{cr}, provided that all other closed-loop system poles are in the left half s-plane. In general, the closer the $L(j\omega)$ locus comes to the $-1 + j0$ point in the Nyquist plot, the more oscillatory is the system response. For this reason, the closeness of the $L(j\omega)$ locus to the -1 point can be used as a measure of the stability margin. Two traditional measures of the stability margin are gain margin and phase margin.

Gain margin and phase margin are usually defined for stable closed-loop systems that are characterized by a minimum phase, loop transfer function $L(s)$. The gain margin is the factor by which the open-loop gain of a stable closed-loop system can be increased before the system goes unstable. The phase margin is the amount of additional phase lag at the gain crossover frequency required to make the stable closed-loop system marginally stable. Thus we have the following definitions:

Gain margin (GM): The gain margin is the reciprocal of the magnitude $|L(j\omega)|$ at the phase crossover frequency ω_ϕ, where the phase of $L(j\omega_\phi)$ reaches $-180°$. The gain margin is given by

$$\text{GM} = \frac{1}{|L(j\omega_\phi)|} \qquad \text{or} \qquad \text{GM(dB)} = -20 \log|L(j\omega_\phi)|$$

Phase margin (PM): The phase margin is defined as the angle between the phase of the loop transfer at the gain crossover frequency ω_g where $|L(j\omega_g)| = 1$ and the angle $-180°$, or $\text{PM} = \angle L(j\omega_g) + 180°$.

The gain and phase margins are shown in Fig. 27.19. Gain and phase margins are stability margins for single-input single-output systems. They cannot apply for multi-input multi-output systems. In addition, they can be a poor indication of stability margin in the face of combined gain and phase variations, as shown in Fig. 27.20. This is due to the fact that gain and phase margins are measures of stability margin in terms of only pure gain and phase variations, but not a combination of both. As a consequence, a system may have good gain and phase margins, but it is close to instability, as indicated in Fig. 27.20. To make up for the insufficiencies of gain and phase margins, a third stability margin, return difference, is used in modern control theory. We will only consider single-input single-output systems.

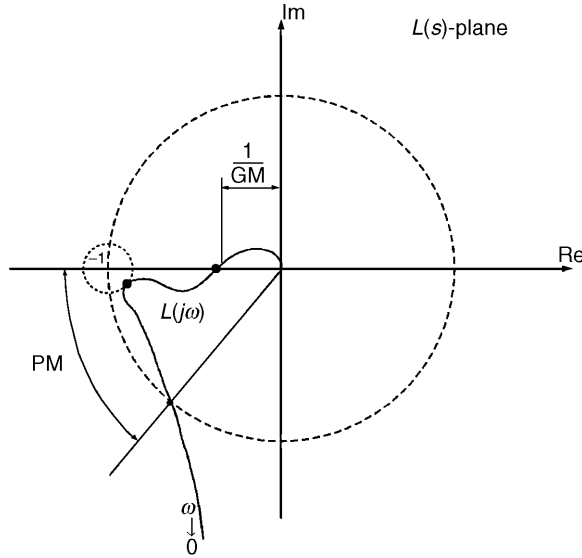



FIGURE 27.20 Insufficiency of gain and phase margins.

Minimum return difference: The minimum return difference is the minimum value of $|1 + L(j\omega)|$, for $0 < \omega < \infty$. It can be seen from Fig. 27.20 that the minimum return difference is the minimum distance from the Nyquist plot to the −1 point. Therefore, the gain and phase margins are special cases of the minimum return difference. The gain margin is directly related to the case when the minimum return difference occurs at the phase crossover frequency, and the phase margin is corresponding to the case that the minimum return difference occurs at the gain crossover frequency.

Although the minimum return difference is a better measure of stability margin than the gain and phase margins, it is seldom used in the classical control theory. This is because the classical control analysis and design is usually carried out using the Bode diagram or the Nichols chart instead of the Nyquist plot. The gain and phase margins are more easily determined from the Bode diagram or the Nichols chart than the Nyquist plot. Despite the fact that the minimum return difference can be easily evaluated from the Nyquist plot, it is difficult to determine the minimum return difference from the Bode plot or the Nichols chart.

We now correlate the phase margin and the damping ratio ζ of an underdamped second-order system. Consider the standard second-order system

$$T(s) = \frac{\omega_n^2}{s^2 + 2\varsigma\omega_n s + \omega_n^2} \tag{27.14}$$

We assume that the transfer function $T(s)$ comes from a unity feedback configuration and can be rewritten as

$$T(s) = \frac{G(s)}{1 + G(s)}$$

where the open-loop transfer function $G(s)$ is given by

$$G(s) = \frac{\omega_n^2}{s(s + 2\varsigma\omega_n)} \tag{27.15}$$

The phase margin occurs at the gain crossover frequency ω_c when $|G(j\omega_c)| = 1$, or

$$\frac{\omega_n^2}{\omega_c(\omega_c^2 + 4\varsigma^2\omega_n^2)^{1/2}} = 1$$

This equation can be rewritten as

$$(\omega_c^2)^2 + 4\varsigma^2\omega_n^2(\omega_c^2) - \omega_n^4 = 0$$

Solving for positive ω_c, we obtain

$$\frac{\omega_c^2}{\omega_n^2} = (4\varsigma^4 + 1)^{1/2} - 2\varsigma^2$$

Substituting $s = j\omega_c$ into Eq. (27.15), the phase margin for the system is

$$\begin{aligned} \text{PM} &= 180° + \angle G(j\omega_c) \\ &= 180° - 90° - \tan^{-1}\left(\frac{\omega_c}{2\varsigma\omega_n}\right) \\ &= 90° - \tan^{-1}\left(\frac{1}{2\varsigma}[(4\varsigma^4 + 1)^{1/2} - 2\varsigma^2]^{1/2}\right) \\ &= \tan^{-1}\left(2\varsigma\left[\frac{1}{(4\varsigma^4 + 1)^{1/2} - 2\varsigma^2}\right]^{1/2}\right) \end{aligned} \tag{27.16}$$

Equation (27.16) relates the damping ratio of the standard second-order system (27.14) to the phase margin of its corresponding open-loop system (27.15) in a unity feedback configuration. This equation provides a correlation between the frequency response and the time response. A plot of ζ versus PM is shown in Fig. 27.21. The curve of ζ versus PM can be approximated by a dashed line in Fig. 27.21. The linear approximation can be expressed as

$$\varsigma = 0.01 \text{ PM} \tag{27.17}$$

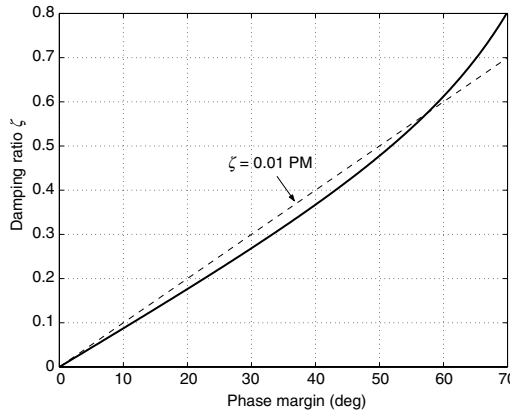



FIGURE 27.21 Damping ratio vs. phase margin for a second-order system.

This approximation is reasonably accurate for $\zeta \leq 0.7$ and is useful in relating the frequency response to the transient performance of a second-order system. Equation (27.17) may also be used for higher-order systems if the system can be assumed to have a pair of dominant underdamped complex poles.

References

1. Dorf, R.C., and Bishop, R.H., *Modern Control Systems,* 9th ed., Prentice-Hall, 2001.
2. Ogata, K., *Modern Control Engineering,* 2nd ed., Prentice-Hall, 1990.
3. Kuo, B.C., *Control Systems,* 7th ed., Prentice-Hall, 1995.
4. Franklin, G.F., Powell, J.D., and Emami-Naeini, A., *Feedback Control of Dynamic Systems,* 3rd ed., Addison-Wesley, 1994.
5. Phillips, C.L., and Harbor, R.D., *Feedback Control Systems,* 4th ed., Prentice-Hall, 2000.
6. Wolovich, W.A., *Automatic Control Systems: Basic Analysis and Design,* Harcourt Brace College Publishing, 1994.

28

Kalman Filters as Dynamic System State Observers

Timothy P. Crain II
NASA Johnson Space Center

28.1 The Discrete-Time Linear Kalman Filter

Distilled to its most fundamental elements, the Kalman filter [1] is a predictor-corrector estimation algorithm that uses a dynamic system model to predict state values and a measurement model to correct this prediction. However, the Kalman filter is capable of a great deal more than just state observation in such a manner. By making certain stochastic assumptions, the Kalman filter carries along an internal metric of the statistical confidence of the estimate of individual state elements in the form of a covariance matrix. The essential properties of the Kalman filter are derived from the requirements that the state estimate be

- a linear combination of the previous state estimate and current measurement information
- unbiased with respect to the true state
- and optimal in terms of having minimum variance with respect to the true state.

Starting with these basic requirements an elegant and efficient formulation for the implementation of the Kalman filter may be derived.

The Kalman filter processes a time series of measurements to update the estimate of the system state and utilizes a dynamic model to propagate the state estimate between measurements. The observed measurement is assumed to be a function of the system state and can be represented via

$$\mathbf{Y}(t) = h(\mathbf{X}(t), \boldsymbol{\beta}, t) + \mathbf{v}(t) \tag{28.1}$$

where $\mathbf{Y}(t)$ is an m dimensional observable, h is the nonlinear measurement model, $\mathbf{X}(t)$ is the n dimensional system state, $\boldsymbol{\beta}$ is a vector of modeling parameters, and $\mathbf{v}(t)$ is a random process accounting for measurement noise.

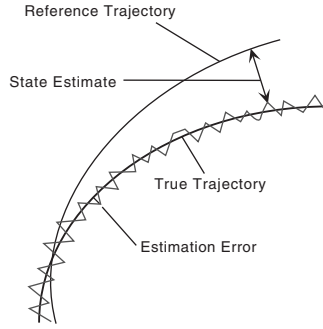



FIGURE 28.1 LKF tracking of a two-dimensional trajectory.

The true dynamic system is described by a general first-order, ordinary differential equation

$$\dot{\mathbf{X}}(t) = f(\mathbf{X}(t), \boldsymbol{\alpha}, t) + \mathbf{w}(t) \tag{28.2}$$

where f is the nonlinear dynamics function that incorporates all significant deterministic effects of the environment, $\boldsymbol{\alpha}$ is a vector of parameters used in the model, and $\mathbf{w}(t)$ is a random process that accounts for the noise present from mismodeling in f or from the quantum uncertainty of the universe, depending on the accuracy of the deterministic model in use.

With these general models available, a linear Kalman filter (LKF) may be derived in a discrete-time formulation. The dynamics and measurement functions are linearized about a known reference state, $\tilde{\mathbf{X}}(t)$, which is related to the true environment state, $\mathbf{X}(t)$, via

$$\tilde{\mathbf{X}}(t) + \mathbf{x}(t) = \mathbf{X}(t) \tag{28.3}$$

The LKF state estimate is related to the true difference by

$$\hat{\mathbf{x}}_k^{(\pm)} = \mathbf{x}_k + \delta\mathbf{x}_k^{(\pm)} \tag{28.4}$$

where the "^" denotes the state estimate (or filter state), $\delta\mathbf{x}_k^{(\pm)}$ is the estimation error, and "±" indicates whether the estimate and error are evaluated instantaneously before (–) or after (+) measurement update at discrete time t_k.

It is important to emphasize that the LKF filter state is the estimate of the difference between the environment and the reference state. The LKF mode of operation will therefore carry along a reference state and the filter state between measurement updates. Only the filter state is at the time of measurement update. Figure 28.1 illustrates the generalized relationship between the true, reference, and filter states in an LKF estimating a two-dimensional trajectory.

Linearization of Dynamic and Measurement System Models

The dynamics and measurement functions may be linearized about the known reference state, $\tilde{\mathbf{X}}(t)$, according to

$$f(\mathbf{X}, \boldsymbol{\alpha}, t) \simeq f(\tilde{\mathbf{X}}(t), \boldsymbol{\alpha}, t) + \mathbf{F}(\tilde{\mathbf{X}}(t), \boldsymbol{\alpha}, t)\mathbf{x}(t) + \mathbf{w}(t) \tag{28.5}$$

$$h(\mathbf{X}, \boldsymbol{\alpha}, t) \simeq h(\tilde{\mathbf{X}}(t), \boldsymbol{\beta}, t) + \mathbf{H}(\tilde{\mathbf{X}}(t), \boldsymbol{\beta}, t)\mathbf{x}(t) + \mathbf{v}(t) \tag{28.6}$$

where $\mathbf{F}$ is the dynamics partial derivative matrix and $\mathbf{H}$ is the measurement partial derivative matrix defined by

$$\mathbf{F}(\tilde{\mathbf{X}}(t), \boldsymbol{\alpha}, t) = \left.\frac{\partial f}{\partial \mathbf{X}}\right|_{\mathbf{X}=\tilde{\mathbf{X}}} \tag{28.7}$$

$$\mathbf{H}(\tilde{\mathbf{X}}(t), \boldsymbol{\beta}, t) = \left.\frac{\partial h}{\partial \mathbf{X}}\right|_{\mathbf{X}=\tilde{\mathbf{X}}} \tag{28.8}$$

and $\mathbf{x}(t)$ is the true state to be estimated representing the difference between the environment and reference states

$$\mathbf{x}(t) = \mathbf{X}(t) - \tilde{\mathbf{X}}(t) \tag{28.9}$$

After linearizing the dynamic and measurement models, the effect of neglecting the higher order terms is assumed to be included in the random processes $\mathbf{w}(t)$ and $\mathbf{v}(t)$. The linearization is an acceptable approximation if $\mathbf{x}(t)$ is sufficiently small.

The reference and filter states are propagated according to the discrete-time linear relationship

$$\tilde{\mathbf{X}}_{k+1} = \boldsymbol{\Phi}(t_{k+1}, t_k)\tilde{\mathbf{X}}_k \tag{28.10}$$

$$\hat{\mathbf{x}}_{k+1}^{(-)} = \boldsymbol{\Phi}(t_{k+1}, t_k)\hat{\mathbf{x}}_k^{(-)} \tag{28.11}$$

where $\boldsymbol{\Phi}(t_{k+1}, t_k)$ is the state transition matrix from time t_k to time t_{k+1} and has the following properties:

$$\begin{aligned} \boldsymbol{\Phi}(t_k, t_k) &= \mathbf{I} \\ \dot{\boldsymbol{\Phi}}(t_{k+1}, t_k) &= \mathbf{F}(\tilde{\mathbf{X}}(t), \boldsymbol{\alpha}, t)\boldsymbol{\Phi}(t_{k+1}, t_k) \\ \boldsymbol{\Phi}(t_{k+2}, t_k) &= \boldsymbol{\Phi}(t_{k+2}, t_{k+1})\boldsymbol{\Phi}(t_{k+1}, t_k) \end{aligned} \tag{28.12}$$

Note that the system dynamics are now incorporated into the propagation of the reference and filter states by the integration of the dynamics partial derivative in Eq. (28.13).

Mathematically, the true difference state is propagated in a similar fashion with the addition of a process noise random value

$$\mathbf{x}_{k+1} = \boldsymbol{\Phi}(t_{k+1}, t_k)\mathbf{x}_k + \mathbf{w}_k \tag{28.13}$$

In general, it is not required that the reference dynamic model be exactly the same as the truth dynamics or that the modeling parameter $\boldsymbol{\alpha}$ be equivalent to the true modeling vector. This notation is left in place to simplify the derivation of the Kalman filter formulation. A number of innovative approaches have been developed for adapting reference model parameters to improve fidelity with the unknown real-world system model [2–6] and can be used to enhance filter operation.

The LKF also requires a linearized measurement, $\mathbf{y}_k = \mathbf{Y}_k - \tilde{\mathbf{Y}}_k$, modeled by

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{x}_k + \mathbf{v}_k \tag{28.14}$$

For the development of the Kalman filters presented here, the random contributions $\mathbf{v}_k$ and $\mathbf{w}_k$ are assumed to be discrete realizations of the continuous zero mean Gaussian process in Eqs. (28.1) and (28.2) and are defined by

$$E\left[\mathbf{v}_k\mathbf{v}_i^T\right] = \mathbf{R}_k\delta_{ki} \tag{28.15}$$

$$E\left[\mathbf{w}_k\mathbf{w}_i^T\right] = \mathbf{Q}_k\delta_{ki} \tag{28.16}$$

Generally, it is assumed that the process noise and measurement noise sequences are uncorrelated so that

$$E\left[\mathbf{w}_k\mathbf{v}_i^T\right] = 0 \quad \forall k, \forall i \tag{28.17}$$

However, the Kalman filter can be configured to operate in systems where this assumption does not apply [7].

Linear Kalman Filter Error Covariance Propagation

The propagation of the filter and reference states in the LKF were outlined in the previous section in Eqs. (28.11) and (28.13). However, all Kalman filter formulations must also propagate a confidence metric of the state estimate in the form of a state error covariance matrix. The state error covariance, **P**, is defined as the expectation of the outer product of the estimation error vector

$$\mathbf{P}_k^{(\pm)} = E\left[\delta\mathbf{x}_k^{\pm}\delta\mathbf{x}_k^{\pm T}\right] \tag{28.18}$$

The state error covariance matrix is $n \times n$ and symmetric, and must remain positive definite to retain filter stability. The mechanism for propagating the covariance can be derived by taking the covariance just before measurement update at time t_{k+1}

$$\mathbf{P}_{k+1}^{(-)} = E\left[\delta\mathbf{x}_{k+1}^{(-)}\delta\mathbf{x}_{k+1}^{(-)T}\right] \tag{28.19}$$

and substituting the estimation error and propagation definitions in Eqs. (28.4), (28.11), and (28.13) to yield

$$\mathbf{P}_{k+1}^{(-)} = E\left[\mathbf{\Phi}(t_{k+1}, t_k)\left(\hat{\mathbf{x}}_k^{(+)} - \mathbf{x}_k\right)\left(\hat{\mathbf{x}}_k^{(+)} - \mathbf{x}_k\right)^T\mathbf{\Phi}(t_{k+1}, t_k)^T + \mathbf{w}_k\mathbf{w}_k^T\right] \tag{28.20}$$

Utilizing the definitions of process noise covariance in Eq. (28.16) and state error covariance in Eq. (28.18) the propagation equation reduces to

$$\mathbf{P}_{k+1}^{(-)} = \mathbf{\Phi}(t_{k+1}, t_k)\mathbf{P}_k^{(+)}\mathbf{\Phi}(t_{k+1}, t_k)^T + \mathbf{Q}_k \tag{28.21}$$

The propagation equation can be interpreted as the sum of the mapping of the previous post-update error covariance through the system dynamics and the system process noise induced uncertainty. Thus, process noise acts to increase the state error covariance between measurement updates.

Linear Kalman Filter Update

The linear Kalman filter (LKF) seeks an unbiased, minimum variance solution for the difference state, $\mathbf{x}_k$, by combining previous state information with available measurements. The state estimate after measurement update is therefore assumed to be a linear combination of the pre-update state and the linearized measurement information

$$\hat{\mathbf{x}}_k^{(+)} = \mathbf{K}_k^*\hat{\mathbf{x}}_k^{(-)} + \mathbf{K}_k\mathbf{z}_k \tag{28.22}$$

Substituting Eqs. (28.4) and (28.14) into Eq. (28.22) and solving for the estimation error yields

$$\delta\mathbf{x}_k^{(+)} = (\mathbf{K}_k^* + \mathbf{K}_k\mathbf{H}_k - \mathbf{I})\mathbf{x}_k + \mathbf{K}_k^*\delta\mathbf{x}_k^{(-)} + \mathbf{K}_k\mathbf{v}_k \tag{28.23}$$

By definition $E[\mathbf{v}_k] = \mathbf{0}$ and $E[\delta\mathbf{x}_k^{(-)}] = \mathbf{0}$ by assumption of unbiased estimation. Therefore, the updated state estimation error is unbiased

$$E\left[\delta\mathbf{x}_k^{(+)}\right] = \mathbf{0} \tag{28.24}$$

only if

$$\mathbf{K}_k^* + \mathbf{K}_k\mathbf{H}_k - \mathbf{I} = \mathbf{0} \tag{28.25}$$

Substitution of Eq. (28.25) into Eq. (28.22) results in an expression for the updated state estimate

$$\hat{\mathbf{x}}_k^{(+)} = \hat{\mathbf{x}}_k^{(-)} + \mathbf{K}_k\left(\mathbf{z}_k - \mathbf{H}_k\hat{\mathbf{x}}_k^{(-)}\right) \tag{28.26}$$

with estimation error

$$\delta\mathbf{x}_k^{(+)} = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\delta\mathbf{x}_k^{(-)} + \mathbf{K}_k\mathbf{v}_k \tag{28.27}$$

The post-measurement error covariance in Eq. (28.18) may be expanded to

$$\mathbf{P}_k^{(+)} = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^{(-)}(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)^T + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^T \tag{28.28}$$

by substitution of Eq. (28.27) and applying the conditions of uncorrelated process and measurement noise, zero mean measurement noise, and the definition of the pre-measurement state estimation error covariance. At this point, only the requirement that the Kalman filter be an unbiased estimator has been satisfied, so now we will select the Kalman gain $\mathbf{K}_k$ that delivers the minimum summed variance on the post-measurement state estimation error. In other words, we seek the gain that will minimize

$$J_k = trace\left[\mathbf{P}_k^{(+)}\right] \tag{28.29}$$

The necessary condition for minimality of J_k is that its partial derivative with respect to the Kalman gain is zero. By employing the following relationship

$$\frac{\partial}{\partial \mathbf{A}}[trace(\mathbf{ABA}^T)] = 2\mathbf{AB} \tag{28.30}$$

where $\mathbf{B}$ is a symmetric matrix, on the components of Eq. (28.28) with respect to $\mathbf{K}_k$ results in

$$(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^{(-)}\mathbf{H}_k^T + \mathbf{K}_k\mathbf{R}_k = \mathbf{0} \tag{28.31}$$

The optimal gain (the Kalman gain) is therefore

$$\mathbf{K}_k = \mathbf{P}_k^{(-)}\mathbf{H}_k^T\left[\mathbf{H}_k\mathbf{P}_k^{(-)}\mathbf{H}_k^T + \mathbf{R}_k\right]^{-1} \tag{28.32}$$

which is sometimes written as

$$\mathbf{K}_k = \mathbf{P}_k^{(-)}\mathbf{H}_k^T\mathbf{W}_k^{-1} \tag{28.33}$$

where the term $\mathbf{W}_k$ is referred to as the innovations covariance

$$\mathbf{H}_k\mathbf{P}_k^{(-)}\mathbf{H}_k^T + \mathbf{R}_k \tag{28.34}$$

28.2 Other Kalman Filter Formulations

In addition to the LKF, there are several other formulations of the Kalman filter that may be employed to more closely follow the characteristics of specific state observation scenarios. The LKF may be varied according to the temporal nature of the dynamic and measurement systems to be continuous in dynamics and measurements or continuous in dynamics and discrete in measurements [12]. Also, there are applications when the dynamic system is energetic or the measurement quality is poor that may cause the reference state in the LKF to quickly leave the region of linearity about the environment state. In such systems, the reference state can be updated through addition of the filter state into an implementation known as the extended Kalman filter (EKF). The EKF is highly suited to real-time applications but is nonlinear in the sense that the reference state is essentially reinitialized at the time of each measurement update. Both the continuous–discrete LKF and EKF will be developed in the following sections.

The Continuous–Discrete Linear Kalman Filter

There may quite naturally arise an application where the reference state, filter state, and state error covariance are more suitably propagated in a continuous fashion than through the linear application of the state transition matrix. Also, it is common for the measurement system to deliver discrete-time observations even when the dynamics are best modeled continuously. In such a situation the update mechanization is unchanged from the previous LKF derivation while the propagation between updates is carried out through continuous integration. Without loss of generality, it may be stated that the reference dynamics of a continuous Kalman filter may be represented by

$$\dot{\tilde{\mathbf{X}}}(t) = f(\tilde{\mathbf{X}}(t), \boldsymbol{\alpha}, t) \tag{28.35}$$

Furthermore, by taking time derivatives of the filter state and covariance propagation (Eqs. (28.11) and (28.21)) and substituting in Eq. (28.13) for the derivative of the state transition matrix, the continuous-time filter state and covariance relations are found to be

$$\dot{\hat{\mathbf{x}}}^{(-)}(t) = f(\tilde{\mathbf{X}}(t), \boldsymbol{\alpha}, t) + \mathbf{F}(\tilde{\mathbf{X}}(t), \boldsymbol{\alpha}, t)[\hat{\mathbf{x}}^{(-)}(t) - \tilde{\mathbf{X}}(t)] \tag{28.36}$$

$$\dot{\mathbf{P}}(t) = \mathbf{F}(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{F}^T(t) + \overline{\mathbf{Q}}(t) \tag{28.37}$$

where $\overline{\mathbf{Q}}(t)$ is the spectral density of the dynamic process noise at time t and the explicit functional dependency of $\mathbf{F}$ was dropped for notational convenience. In this mechanization of the LKF, the state transition matrix need not be calculated as the dynamics are included directly via the partial derivative matrix and the reference state, filter state, and error covariance are propagated continuously.

The process and measurement noise representations in this formulation are continuous and discrete for the respective models, and are again assumed to be zero mean processes governed by the continuous dynamic process noise covariance

$$E[\mathbf{w}(t)\mathbf{w}^T(\tau)] = \mathbf{Q}(t)\delta(t-\tau) \tag{28.38}$$

and the discrete measurement noise covariance

$$E\left[\mathbf{v}_k\mathbf{v}_j^T\right] = \mathbf{R}_k\delta_{kj} \tag{28.39}$$

It is also assumed here that the process and measurement noises are uncorrelated so that

$$E\left[\mathbf{w}(t)\mathbf{v}_k^T\right] = 0 \tag{28.40}$$

although the formulation can be modified to accommodate process and measurement noise correlations if necessary [7].

The Continuous–Discrete Extended Kalman Filter

In applications where the reference state may quickly deviate beyond the linear region of the environment state, the reference may be directly updated at the time of measurement update by adding the LKF filter state to the reference in an EKF. The EKF is similar to the LKF, in that measurements are processed to provide an estimate of the difference between the true state and reference state of the spacecraft. Also, the EKF evaluates dynamics and measurement partials with respect to the reference state in a manner similar to the LKF. However, the reference state about which these partials are evaluated is modified through the addition of measurement information

$$\tilde{\mathbf{X}}(t_k)^{(+)} = \tilde{\mathbf{X}}(t_k)^{(-)} + \hat{\mathbf{x}}(t_k) \tag{28.41}$$

The reference state dynamics model used in the EKF formulation is given by Eq. (28.35), but the measurement model is the discrete form given by Eq. (28.1). The filter state representing the estimated difference between the true state and the reference state is only calculated at the time of measurement update via dropping the previous estimate information term from Eq. (28.26):

$$\hat{\mathbf{x}}_k = \mathbf{K}_k \mathbf{Z}_k \tag{28.42}$$

where the innovation is now the actual measurement residual

$$\mathbf{Z}_k = \mathbf{Y}_k - \mathbf{h}\left(\tilde{\mathbf{X}}_k^{(-)}, \boldsymbol{\beta}, t_k\right) \tag{28.43}$$

Therefore, in the EKF there is not a separate filter state that needs to be propagated to the time of the next measurement, as the filter state has been incorporated into the updated reference state.

As before, the error covariance at each measurement is updated by

$$\mathbf{P}_k^{(+)} = (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^{(-)}(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)^T + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^T, \tag{28.44}$$

and the EKF Kalman gain and innovations covariance are analogous to their LKF counterparts

$$\mathbf{K}_k = \mathbf{P}_k^{(-)}\mathbf{H}_k\mathbf{W}_k^{-1} \tag{28.45}$$

$$\mathbf{W}_k = \mathbf{H}_k\mathbf{P}_k^{(-)}\mathbf{H}_k + \mathbf{R}_k. \tag{28.46}$$

The difference between EKF operation and LKF operation is illustrated by revisiting the two-dimensional trajectory illustration in Fig. 28.2. The reference trajectory can now be seen to respond to measurement information availability and tracks the true environment trajectory.

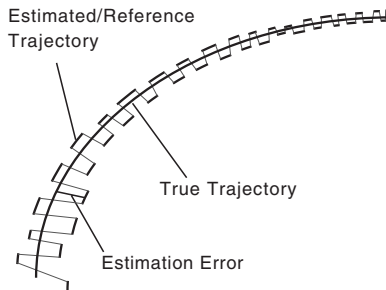



FIGURE 28.2 EKF tracking of a two-dimensional trajectory.

28.3 Formulation Summary and Review

The LKF discrete–discrete formulation was given by the following propagation equations:

$$\begin{aligned}
\tilde{\mathbf{X}}_{k+1} &= \mathbf{\Phi}(t_{k+1}, t_k)\tilde{\mathbf{X}}_k \\
\hat{\mathbf{x}}_{k+1}^{(-)} &= \mathbf{\Phi}(t_{k+1}, t_k)\hat{\mathbf{x}}_k^{(-)} \\
\mathbf{P}_{k+1}^{(-)} &= \mathbf{\Phi}(t_{k+1}, t_k)\mathbf{P}_k^{(-)}\mathbf{\Phi}(t_{k+1}, t_k)^T + \mathbf{Q}_k
\end{aligned}$$

and update equations

$$\begin{aligned}
\hat{\mathbf{x}}_k^{(+)} &= \hat{\mathbf{x}}_k^{(-)} + \mathbf{K}_k\left[\mathbf{z}_k - \mathbf{H}_k\hat{\mathbf{x}}_k^{(-)}\right] \\
\mathbf{P}_k^{(+)} &= (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^{(-)}(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)^T + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^T \\
\mathbf{K}_k &= \mathbf{P}_k^{(-)}\mathbf{H}_k^T\left[\mathbf{H}_k\mathbf{P}_k^{(-)}\mathbf{H}_k^T + \mathbf{R}_k\right]^{-1}
\end{aligned}$$

In the discrete time LFK mechanization, the reference state is unaffected by the incorporation of measurement information into the filter state.

In a slight variation of this approach, the dynamics of the LKF may be made continuous, and the filter state, reference state, and covariance propagated without the use of a state transition matrix.

$$\begin{aligned}
\dot{\tilde{\mathbf{X}}}(t) &= f(\tilde{\mathbf{X}}(t), \boldsymbol{\alpha}, t) \\
\dot{\hat{\mathbf{x}}}^{(-)}(t) &= \mathbf{F}(\tilde{\mathbf{X}}(t), \boldsymbol{\alpha}, t)\hat{\mathbf{x}}^{(-)}(t) \\
\dot{\mathbf{P}}(t) &= \mathbf{F}(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{F}^T(t) + \bar{\mathbf{Q}}(t)
\end{aligned}$$

When the application requires that the reference state be modified to remain in the linear vicinity of the environment state, the EKF continuous–discrete formulation may be appropriate. In the continuous–discrete EKF formulation, the propagation is carried out according to

$$\begin{aligned}
\dot{\tilde{\mathbf{X}}}(t) &= f(\tilde{\mathbf{X}}(t), \boldsymbol{\alpha}, t) \\
\dot{\mathbf{P}}(t) &= \mathbf{F}(t)\mathbf{P}(t) + \mathbf{P}(t)\mathbf{F}^T(t) + \bar{\mathbf{Q}}(t)
\end{aligned}$$

and the measurement update according to

$$\begin{aligned}
\tilde{\mathbf{X}}(t_k)^{(+)} &= \tilde{\mathbf{X}}(t_k)^{(-)} + \hat{\mathbf{x}}(t_k) \\
\hat{\mathbf{x}}_k &= \mathbf{K}_k\left[\mathbf{Y}_k - \mathbf{h}\left(\tilde{\mathbf{X}}_k^{(-)}, \boldsymbol{\beta}, t_k\right)\right] \\
\mathbf{P}_k^{(+)} &= (\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)\mathbf{P}_k^{(-)}(\mathbf{I} - \mathbf{K}_k\mathbf{H}_k)^T + \mathbf{K}_k\mathbf{R}_k\mathbf{K}_k^T \\
\mathbf{K}_k &= \mathbf{P}_k^{(-)}\mathbf{H}_k^T\left[\mathbf{H}_k\mathbf{P}_k^{(-)}\mathbf{H}_k^T + \mathbf{R}_k\right]^{-1}
\end{aligned}$$

The reference state will change with the incorporation of measurement information into the EKF and the partials evaluated along this changing reference.

28.4 Implementation Considerations

It is commonly held among designers of Kalman filters that the implementation of the formulas listed above represent only a portion of the effort required to develop an accurate and robust Kalman filter application. Once the dynamics, measurements, and partial derivatives have been coded, the task remains to tune the noise magnitudes represented in the process noise covariance **Q** and the measurement noise covariance **R**. While the measurement noise can be based in realistic hardware performance specifications, the process noise is often used as a tuning parameter to ensure that the filter operates correctly. This process of tuning the filter crosses over into the area of design and is nearly an art form of such myriad approaches that it is beyond the scope of this work to outline. However, a Kalman filter checklist is provided for the newcomer to the field to reduce the time of the implementation and tuning learning curve:

- Because the linear Kalman filter does not change the reference state in the presence of measurement information, the reference state and partial derivatives for an LKF application may be computed prior to operation. This makes the LKF more amenable to computationally restricted applications or hypothesis testing where differing process noise and measurement noise parameters are being evaluated in parallel [8].
- Process noise serves to keep the filter from becoming overconfident in its estimate (i.e., a covariance with near zero diagonal values) and converging prematurely. Examining the propagation equations for the Kalman filters presented previously, it can easily be seen how the addition of process noise increases the magnitude of the state error covariance between measurements.
- The innovations covariance should ideally converge to describe the variance in the filter measurement residuals. Adaptive techniques have been implemented where the filter noise parameters are tuned according to a metric linking residual statistics with the innovations covariance [5]. In an ideal filter, the innovations covariance should approach the measurement noise covariance as the process noise magnitude approaches zero.
- When multiple measurements are available at the same time, they may be processed as a series of scalar observations as long as they are uncorrelated (i.e., **R** is a diagonal matrix). The effect of processing scalar measurements is that the innovations covariance becomes a scalar, and a numerical division rather than a matrix inversion is required to calculate the Kalman gain.
- Measurement editing may be employed to prevent spurious data from causing filter divergence in a number of ways. One of the most common is to reject measurements when the ratio of the measurement residual squared to the scalar innovations covariance

$$\frac{r_k^2}{W_k} \tag{28.47}$$

  is above a user-defined threshold. The threshold value may either be a constant or may be time varying after long propagation periods to allow for a smooth transition to a steady state innovations covariance.
- The covariance should always be positive definite. If filter divergence is a chronic problem in a particular application, the numerical integrity of the covariance may provide insight into the nature of the divergence. There are also several numerical implementations of the covariance update equation that take advantage of its symmetry and positive definiteness to enhance its stability while reducing computational load [9].
- Process noise may be enhanced by including time correlated states such as first-order Gauss–Markov processes to the filter to account for specific dynamic effects. The biases associated with these processes can be included in the filter state for estimation.

As a final note it should be stressed that the Kalman filter is not the state observer algorithm best suited for all applications. Its strengths lie in light computational requirements and real-time availability

of a state estimate in the presence of accurate measurement information. However, batch estimation techniques such as least-squares estimation may be more appropriate in applications where the dynamic process is modeled to a high degree of fidelity, measurements are not uniformly accurate, and real-time operation is not an issue. A number of quality texts [10–12] have been written on the subject of stochastic estimation in general and specifically Kalman filtering that the the reader is encouraged to pursue for more detailed information.

References

1. Kalman, R. E., "A new approach to linear filtering and prediction problems," *Transactions of the ASME, Ser. D, Journal of Basic Equations,* March 1960, pp. 35–45.
2. Burkhart, P. and Bishop, R., "Adaptive orbit determination for interplanetary spacecraft," *Journal of Guidance, Control, and Dynamics,* Vol. 19, No. 3, 1997, pp. 693–701.
3. Chaer, W., Bishop, R., and Ghosh, J., "Hierarchical adaptive Kalman filtering for interplanetary orbit determination," *IEEE Transactions on Aerospace and Electronic Systems,* Vol. 34, No. 3, 1998, pp. 1–14.
4. Crain, T. and Bishop, R., "The mixture-of-experts gating network: integration into the ARTSN extended Kalman filter," Technical Memorandum CSR-TM-99-01, Center for Space Research, March 1999.
5. Ely, T., Bishop, R., and Crain, T., "Adaptive interplanetary navigation using genetic algorithms," *The Journal of Astronautical Sciences,* 2000, Accepted for Publication.
6. Crain, T. and Bishop, R., "Unmodeled impulse detection and identification during Mars pathfinder cruise," Technical Memorandum CSR-TM-00-01, Center for Space Research, March 2000.
7. Chaer, W. and Bishop, R., "Adaptive Kalman filtering with genetic algorithms," *Advances in the Astronautical Sciences,* edited by R. Proulx, J. Liu, P. Siedelmann, and S. Alfano, Vol. 89, Univelt, San Diego, CA, 1995, pp. 141–156, Pt. 1.
8. Gholson, N. and Moose, R., "Maneuvering target tracking using adaptive state estimation," *IEEE Transactions on Aerospace and Electronic Systems,* Vol. 13, No. 3, May 1997, pp. 310–317.
9. Bierman, G., *Factorization Methods for Discrete Sequential Estimation,* Academic Press, 1977.
10. Brown, R. G. and Huang, P. Y. C., *Introduction to Random Signals and Applied Kalman Filtering,* John Wiley and Sons, 1992.
11. Lewis, F., *Applied Optimal Control and Estimation,* Prentice-Hall, Englewood Clifis, NJ, 1992.
12. Gelb, A., *Applied Optimal Estimation,* The M.I.T. Press, Cambridge, MA, 1974.

29

Digital Signal Processing for Mechatronic Applications

Bonnie S. Heck
Georgia Institute of Technology

Thomas R. Kurfess
Georgia Institute of Technology

29.1 Introduction

Most engineers work in the world of mechatronics as there are relatively few systems that are purely mechanical or electronic. There are a variety of means by which electrical systems augment mechanical systems and vise versa. For example, most microprocessors found in a computer today have some sort of heat sink and perhaps a fan attached to them to keep them within their operational temperature zone. Electrical systems are widely employed to monitor and control a wide variety of mechanical systems. With the advent of inexpensive digital processing chips, digital filtering and digital control for mechanical systems is becoming commonplace. Examples of this can be seen in every automobile and most household appliances. For example, sensor signals used in monitoring and controlling of mechanical systems require some form of signal processing. This signal processing can range from simply "cleaning-up" the signal using a low pass filter to more advanced analyses such as torque and power monitoring in a DC servo motor. This chapter presents a brief overview of digital signal processing methods suitable for mechanical systems. Since this chapter is limited in space, it does not give any derivation or details of analysis. For a more detailed discussion, see references [1,2].

29.2 Signal Processing Fundamentals

A few fundamental concepts on signal processing must be introduced before a discussion of filtering or control can be undertaken.

Continuous-Time Signals

Laplace transforms are used for system analysis of continuous-time systems, solving for system response, and control design. The single-sided Laplace transform of a continuous-time signal, $x(t)$, is given by

$$X(s) = \int_0^{\infty} x(t)e^{-st}\,dt$$

A transfer function of a linear system, $H(s)$, can be found as the ratio of the Laplace transforms of the output over that of the input (with zero initial conditions).

The Fourier transform is used to determine the frequency content of a signal. The Fourier transform of $x(t)$ is given by

$$X(\omega) = \int_{-\infty}^{\infty} x(t)e^{-j\omega t}\,dt \tag{29.1}$$

where ω is in units of radians per second. Notice that when $x(t) = 0$ for $t \le 0$, the Laplace transform is equivalent to the Fourier transform by setting $s = j\omega$. (It should be noted that there are some additional convergence considerations for the Fourier transform.) The frequency response of a system is defined as the ratio of the Fourier transforms of the output over that of the input. Equivalently, it can be found from the transfer function as $H(\omega) \equiv H(j\omega) = H(s)|_{s=j\omega}$. For simplicity of notation, the j is usually not shown in the argument list, giving rise to the notation $H(\omega)$ to represent the frequency response. The bandwidth of a system is defined as the frequency at which $H(\omega) = 0.707H(0)$.

Discrete-Time Signals

The z-transform is useful for solving a difference equation and for performing system analysis. The z-transform of a discrete-time signal, $x[n]$, is defined as

$$X(z) = \sum_{n=-\infty}^{\infty} x[n]z^{-n}$$

The discrete-time Fourier transform (DTFT) is used to determine the frequency content of a signal. The DTFT and the inverse DTFT of a signal are defined by

$$X(\Omega) = \sum_{n=-\infty}^{\infty} x[n]e^{-j\Omega n} \tag{29.2}$$

and

$$x[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(\Omega)e^{j\Omega n}\,d\Omega \tag{29.3}$$

Note that the DTFT can be derived from the z-transform by setting $z = e^{j\Omega}$. (Again, there are some assumptions on convergence in this derivation.) Since the DTFT is periodic with period 2π, it is typically displayed over the range $[-\pi, \pi]$ or $[0, 2\pi]$, where the frequencies of general interest are from $\Omega = 0$ (low frequency) to $\Omega = \pi$ (high frequency). The frequency response of a discrete-time system can be found as the ratio of the DTFT of the output signal over that of the input signal. Alternatively, it can be found from the transfer function as $H(\Omega) \equiv H(e^{j\Omega}) = H(z)|_{z=e^{j\Omega}}$. The notation $H(\Omega)$ is preferred over $H(e^{j\Omega})$ for its simplicity. As in the continuous-time case, the bandwidth is defined as the frequency at which $H(\Omega) = 0.707H(0)$.

While the DTFT is continuous with respect to the frequency variable Ω, the discrete Fourier transform (DFT) contains points that are discrete with respect to a parameter k. Consider a finite duration sequence $x[n]$, where $x[n] = 0$ for $n < 0$ and for $n \geq N$. The DFT of $x[n]$ and the inverse DFT are defined as

$$X_k = \sum_{n=0}^{N-1} x[n] e^{-j2\pi nk/N}, \quad k = 0, 1, \ldots, N-1 \tag{29.4}$$

and

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{j2\pi nk/N}, \quad n = 0, 1, \ldots, N-1$$

Note that the DFT is a discretized version of the DTFT where $X_k = X(\Omega)|_{\Omega = 2\pi k/N}$ over the range $\Omega = 0$ to $\Omega = 2\pi$. Calculating a closed-form solution for the DTFT can be done only for simple signals such as a square pulse or a triangular pulse. Therefore, the DFT is generally used as a numerical method to calculate the DTFT at discrete points in frequency in the range $0 \leq \Omega \leq 2\pi$. In particular, to obtain a plot of the DTFT, plot X_k versus k where k is scaled by $2\pi/N$. For an arbitrary signal, such as obtained from measurements of a physical device, computing the DFT instead of the DTFT is the preferred method to find the frequency content of the signal. To get more resolution in plotting a DTFT from the points calculated by a DFT, zeros can be added to the end of the sequence so that the value of N is increased.

Suppose a time domain signal is not finite in duration, so that there is no value of N such that $x[n] = 0$ for $n \geq N$. In order to perform the DFT, the signal must be truncated. There are two cases to be considered: the case where $x[n]$ is decaying to zero and the case where $x[n]$ has periodic components. The case when $x[n]$ decays to zero is handled by choosing N to be large enough so that the signal is negligible beyond that value. The resulting DFT is an approximation (not a discretized version) of the DTFT. If the signal is periodic, the DTFT cannot be computed numerically since the resulting DTFT would have impulses in it. However, the frequencies present in the signal could still be determined if the value of N used for the truncation is chosen so that the truncated signal goes through an integer number of cycles. If this not done, the resulting DFT will have leakage in the frequency plot when compared to the DTFT of the true signal. For example, consider a signal $x[n] = \cos(0.4\pi n)$. This is periodic with period $n = 5$ and has DTFT given by $X(\Omega) = \pi[\delta(\Omega + 0.4\pi) + \delta(\Omega - 0.4\pi)]$ for $-\pi \leq \Omega \leq \pi$. All the frequency content is located at $\Omega = 0.4\pi$ and $\Omega = -0.4\pi$. Since the DTFT is periodic with 2π, there is also an impulse at $\Omega = 2\pi - 0.4\pi$. The DFT is computed for two truncations of the signal, one at $N = 20$ (four complete cycles) and the other at $N = 22$. The DFT for $N = 20$ is plotted in Fig. 29.1(a) where the independent variables k are scaled by $2\pi/N$ for the plot. This plot shows zero frequency content except at $\Omega = 0.4\pi$ (=1.2566) and $\Omega = 2\pi - 0.4\pi$ (=5.0265), giving the correct location of the impulses in the DTFT. Similarly, the DFT for $N = 22$ is plotted in Fig. 29.1(b), notice the resulting leakage in the frequency characteristics.

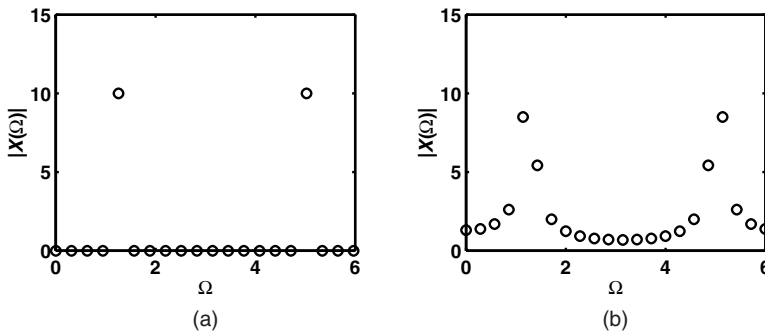



FIGURE 29.1 DFT of periodic signal (a) truncated after 4 complete cycles and (b) truncated after 4.4 cycles.

If a signal has periodic content, but is not periodic, such as $x[n] = \cos(0.5\pi n) + \cos(0.2n)$, then leakage cannot be avoided by a selection of N. An alternate means of reducing leakage is to first taper the signal to zero at the initial and end points of the sequence prior to computing the DFT. This process, known as *windowing* the data, is accomplished by multiplying $x[n]$ by a window function $w[n]$ and then performing the DFT on the product $x[n]w[n]$. Three common windows are the rectangular window, which is a sharp truncation, the Hanning window, and the Hamming window [1].

Rectangular Window:

$$w[n] = 1, \quad 0 \le n \le N-1$$

Hanning Window:

$$w[n] = \frac{1}{2}\left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right), \quad 0 \le n \le N-1$$

Hamming Window:

$$w[n] = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1$$

If the value of N in a DFT is a power of 2, there is a fast method to compute the DFT called the fast Fourier transform (FFT). If the value of N is not a power of 2, zeros can be padded to the end of the signal in order to use the FFT. This does not affect the accuracy of the result, but it does improve the resolution of the resulting plot when the DFT (or FFT) is used to compute the DTFT. In many cases, the expression used in Eq. (29.4) suffices to compute the DFT since the added computational power of today's processors lessens the need for the numerical efficiency of the FFT. The details of the algorithm for the FFT are beyond the scope of this handbook. See [1] or [2] for details.

29.3 Continuous-Time to Discrete-Time Mappings

While most physical systems operate in continuous-time, computers operate in discrete-time. Therefore, in order to use computers to process measurements taken from continuous-time systems, there must be ways of mapping between the continuous-time world to the discrete-time world.

Discretization

Before an analog signal can be analyzed using digital techniques, it must be discretized (that is, converted into a discrete-time signal). The ideal method for discretization is *sampling*, where the values of the signal are determined at discrete points in time. Generally, the signal is sampled at a fixed rate known as the *sampling period*. The sampling rate (in hertz) is the inverse of the sampling period. Figure 29.2 depicts a 1 Hz signal that has been sampled at two rates. The dark points are sampled at 15 ms intervals, while the lighter points are sampled at 250 ms intervals. From Fig. 29.2, the waveform approximation clearly degrades as the sampling frequency is reduced and approaches the signal frequency. In fact, it can be shown that a signal must be sampled at a frequency that is higher than twice its maximum frequency content. This is known as the Nyquist Sampling Theorem. For example, if the signal in Fig. 29.2 is sampled at 0.5 Hz, it is possible for every sample to have a value of 0 as at 0, 500, 1000 ms, etc… the value of the signal is 0. The erroneous interpretation of the signal due to a sample frequency that is too low is known as aliasing. There are two means by which the Nyquist Sampling Theorem can be satisfied. The first is by employing a sample frequency that is more than twice the highest frequency content of the signal being sampled. This value frequency is known as the Nyquist frequency. As one never is sure of the actual frequency content of a real signal, a low pass filter may be used to ensure that a signal does not possess frequencies above a certain cut-off level. Such a filter is commonly

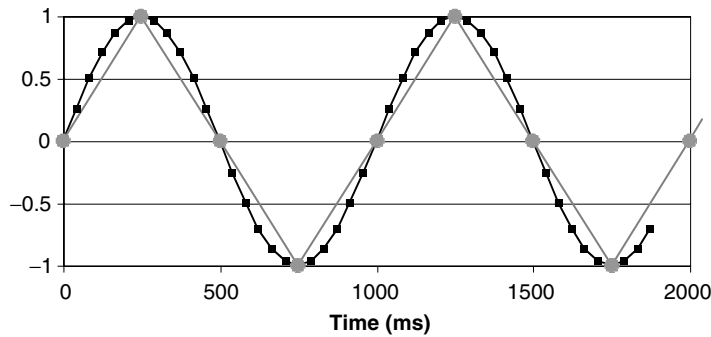



FIGURE 29.2 A 1-Hz signal.

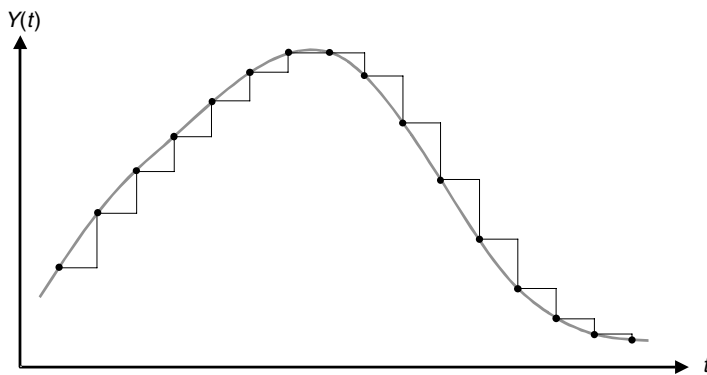



FIGURE 29.3 A signal sampled and reconstructed using a zero order hold (ZOH).

called an anti-aliasing filter. This is the second and most practical method that is used to satisfy the Nyquist Sampling Theorem. Thus, a combination of a well-designed anti-aliasing filter as well as a sampling frequency that is well above the cut-off frequency of the filter will ensure that the Nyquist Sampling Theorem is satisfied.

There are two important points that should be noted when using an anti-aliasing filter. First, it is important that the anti-aliasing filter be used before the signal is sampled as sampling is what causes aliasing. Basically, this requires that the anti-aliasing filter is implemented using an analog filter prior to the signal being digitized. Once a signal has been aliased during sampling, it cannot be corrected using digital filtering. The second point is that, in practice, the cutoff frequency of the anti-aliasing filter should be a factor of 5–10 below the value of the Nyquist frequency. It should be noted that an anti-aliasing filter adds phase lag to the measurement, which might deteriorate stability and performance in a feedback loop unless the bandwidth of the anti-aliasing filter is much higher than that of the closed loop system. Commercially available devices that perform sampling are analog-to-digital converters (ADCs), and the anti-aliasing filter is used before this device.

The converse of sampling is reconstruction where a discrete-time signal is converted into a continuous-time signal. The Nyquist sampling rate ensures that if a continuous-time signal is sampled at a rate that is at least twice the highest frequency component in the signal, then the continuous-time signal can be reconstructed exactly from the samples. However, this theorem assumes that an ideal reconstruction process is available, which is not practical. The most common practical means to reconstruct a signal is a zero-order hold (ZOH). The ZOH assumes that the value of the signal is constant between samples. This approximation is quite reasonable if the sampled signal does not change substantially between individual samples. Figure 29.3 is an example of a signal and its ZOH representation. The gray, smooth line represents the original analog signal. The black points along the signal indicate sample values of the signal. Each black

point is connected to the next via a horizontal then vertical straight line. The horizontal line is representative of the ZOH assumption that the value of the signal remains constant between samples. The vertical line is the reality that the signal does not remain constant over the sample period. As the time between sample points is increased, the accuracy of the ZOH decreases. Conversely, as the sample period is decreased, the accuracy of the ZOH is improved. Commercially available devices that perform reconstruction are digital-to-analog converters (DACs), which generally use the ZOH method.

s-Plane to *z*-Plane Mappings

One method to relate the *s*-plane to the *z*-plane is to derive a continuous-time mathematical representation of the sampled signal for $x(t)$ and compute its Laplace transform. The resulting Laplace transform of the sampled signal can be related to the *z*-transform of $x[n]$ by setting $z = e^{sT}$ where T is the sampling period. The relationship $z = e^{sT}$ is commonly termed the *exact mapping* between the *z*-plane and the *s*-plane. (For details of this derivation, see [1]). For example, a digital representation, $H_d(z)$, of a continuous-time system, $H(s)$, can be obtained using this mapping $H_d(z) = H(s)|_{z=e^{sT}}$. However, this mapping results in a nonrational function for $H_d(z)$.

Approximate mappings between the *s*-plane and the *z*-plane are commonly used that do result in a rational function for $H_d(z)$. Three such mappings are the bilinear transformation, forward transformation, and backward transformation.

Bilinear transformation:

$$s = \frac{2(z-1)}{T(z+1)}$$

Forward transformation:

$$s = \frac{1}{T}(z-1)$$

Backward transformation:

$$s = \frac{1}{Tz}(z-1)$$

The bilinear transformation (also known as Tustin's rule or the trapezoidal rule) is the most accurate of these mappings. It maps the entire left-hand side of the *s*-plane into the unit circle of the *z*-plane, so that it preserves stability. Consider a first-order example of $H(s) = 1/(s + 2)$. The discrete-time representation of this transfer function is

$$H_d(z) = H(s)\Big|_{s=\frac{2(z-1)}{T(z+1)}} = \frac{T(z+1)}{2z-2+2T}$$

Note that the resulting transfer function is rational in *z*.

An alternate method of mapping transfer functions between the continuous-time and the discrete-time domains is the response-matching mapping.

Response-matching: Suppose $x(t)$ is the input to a system $H(s)$ with the resulting output $y(t)$. Let $x[n]$ and $y[n]$ be the sampled versions of $x(t)$ and $y(t)$. Then, $H_d(z)$ is found from the ratio of *z*-transforms of $x[n]$ and $y[n]$. The most common response matching is step-response matching where $x(t)$ is a step function, $x(t) = 1$ for $t \geq 0$ and $x(t) = 0$ for $t < 0$. An expression for $H_d(z)$ is found from the following operation:

$$H_d(z) = (1-z^{-1})Z\left[\frac{H(s)}{s}\right]$$

where $Z[H(s)/s]$ represents the z-transform of the sampled version of the step response of the continuous-time system. The form for a generic first-order system is given below:

$$H(s) = \frac{a}{s+a} \Leftrightarrow H_d(z) = \frac{(1-e^{-aT})z^{-1}}{1-e^{-aT}z^{-1}}$$

The response matching method (especially with step inputs) is commonly used to map a continuous-time plant to discrete-time when designing a digital controller in the discrete-domain. Since most digital controllers are implemented using a ZOH on the output of the digital controller, the plant sees a stepped signal, one that looks like a sum of delayed step signals. Therefore, the step-response matching method is the most accurate way to map a plant that has a ZOH on its input.

Frequency Domain Mappings

The continuous-time Fourier transform can be related to the DTFT through the expression:

$$X(\omega) = TX(\Omega)|_{\Omega=\omega T} \quad \text{for } -\pi \le \Omega \le \pi$$

where $X(\omega)$ is defined in Eq. (29.1) and represents the continuous-time Fourier transform of $x(t)$, while $X(\Omega)$ is defined in Eq. (29.2) and represents the DTFT of the sampled signal $x[n]$. This mapping is very useful for computing the Fourier transform of measured data. In particular, suppose a continuous-time signal is measured by sampling it through an ADC and storing it as a discrete-time sequence. If the signal $x(t)$ is finite in duration, the DTFT of $x[n]$ can be computed at discrete points in frequency by using the DFT, X_k, as given in Eq. (29.4). Using the relationships $\omega = \Omega/T$, $\Omega = 2\pi k/N$, and $X_k = X(\Omega)|_{\Omega=2\pi k/N}$, where N is the length of the sequence for $x[n]$ and T is the sampling period, gives the relationship:

$$X(\omega)|_{\omega=2\pi k/NT} = TX_k \quad \text{for } 0 \le \omega \le \omega_s/2,\ \ 0 \le k \le (N-1)/2$$

where $\omega_s = 2\pi/T$ is the sampling frequency in radian per second. Accuracy can be improved by decreasing the sampling period T, and the resolution in the plot can be increased by increasing NT.

If the signal $x(t)$ is not finite in duration, it must be truncated in order to use this numerical method to calculate the continuous-time Fourier transform. As discussed in the section "Discrete-Time Signals," if the sampled signal $x[n]$ decays to zero, choose the number of sampled points N to be large enough so that $x(t)$ is negligible beyond that value. If the sampled signal $x[n]$ is periodic, choose the sampling period T and the number of points N such that the sampled signal $x[n]$ goes through an integer number of cycles. For example, consider the signal $x(t) = \cos(\pi t)$. If the sampling period is chosen as $T = 0.4$ s, the discretized signal would be $x[n] = x(nT) = \cos(0.4\pi n)$, which is the same signal analyzed in the section "Discrete-Time Signals." Choosing $N = 5$, 10, 15, etc. would yield correct results in the DFT while any other value would result in leakage. If the signal $x(t)$ has periodic content, but does not appear to be periodic, then use a windowing function as discussed in the section "Discrete-Time Signals" to reduce leakage when computing the DFT.

Note that the DFT can also be used to determine the Fourier coefficients of periodic signals. Consider a Fourier series in the form

$$x(t) = \sum_{k=-\infty}^{\infty} c_k e^{j\omega k t}$$

Sample the signal $x(t)$ by first making sure that the sampled signal goes through an integer number of cycles. The coefficients c_k for $k = 0,\ldots,(N-1)/2$ can be found as $c_k = X_k/N$ where X_k is found from the DFT. The rest of the coefficients are obtained from $c_{-k} = c_k^*$.

Since the frequency response of a system is the ratio of the Fourier transforms of the output over the input, a mapping between a continuous-time system, $H(\omega)$, and a corresponding discrete-time system, $H_d(\Omega)$, can be derived from the previous mapping as

$$H(\omega) = H_d(\Omega)|_{\Omega=\omega T} \quad \text{for } -\pi \leq \Omega \leq \pi$$

This mapping is useful for the design of both digital filters and digital controllers.

29.4 Digital Filter Design

The frequency response function of a discrete-time system describes how the system processes input signals of different frequencies. Consider an input signal $x[n] = A\cos(\Omega_0 n)$ to a system with frequency response $H(\Omega)$ where $0 \leq \Omega_0 \leq 2\pi$. The corresponding output is given by

$$y[n] = |H(\Omega_0)|\cos(\Omega_0 n + \angle H(\Omega_0))$$

For aperiodic signals, the filtering property of Fourier transforms gives the relationship:

$$Y(\Omega) = H(\Omega)X(\Omega)$$

Thus, if $|H(\Omega)|$ is small over a certain range of frequencies, then input signals with frequency content in that range are attenuated as they pass through the system.

It is often convenient to filter continuous-time signals through a digital filter as shown in Fig. 29.4. The analog-to-digital converter (ADC) samples the continuous-time signal creating a sequence of discrete-time signals for processing by the computer or digital signal processing board. The filtered signal can be stored digitally for further study or it can be sent through a digital-to-analog converter (DAC). The digital filter can be implemented in software by a recursive equation obtained from the difference equation. Consider a digital filter with transfer function:

$$H(z) = \frac{b_1 z^N + b_2 z^{N-1} + \cdots + b_{N+1}}{a_1 z^N + a_2 z^{N-1} + \cdots + a_{N+1}} = \frac{b_1 + b_2 z^{-1} + \cdots + b_{N+1} z^{-N}}{a_1 + a_2 z^{-1} + \cdots + a_{N+1} z^{-N}}$$

The recursion used to calculate the current value of the output $y[n]$ is given by the difference equation:

$$y[n] = \frac{1}{a_1}(b_1 x[n] + b_2 x[n-1] + \cdots + b_{N+1} x[n-N] - a_2 y[n-1] - \cdots - a_{N+1} y[n-N]) \quad (29.5)$$

Notice that the past values of y and x must be stored for use in the recursion.

Now consider the impulse response of a digital filter, where $y[n]$ is calculated for an input $x[n]$ equal to an impulse (i.e., $\delta[n] = 1$ when $n = 0$ and $\delta[n] = 0$ otherwise). The recursion shown above results in a response for $y[n]$ that has infinite duration (i.e., there is no value of M so that $y[n] = 0$ for all $n > M$). This type of filter is called an *infinite impulse response* (*IIR*) *filter.*

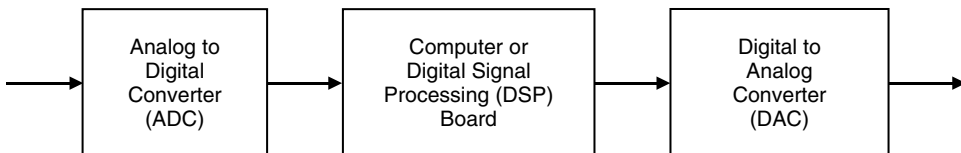



FIGURE 29.4 Configuration for standard digital signal processing hardware.

Now consider the case where the coefficients of the filter $a_m = 0$ for $m > 1$. The resulting expression for $y[n]$ from Eq. (29.5) would no longer be recursive since it would depend only on present and past values of x, and not on past values of y. As a result, the impulse response would have duration N. This type of filter is called *a finite impulse response* (*FIR*) *filter*. FIR filters are sometimes preferred over IIR filters since they have linear phase in the frequency response. Linear phase means that the angle of the frequency response is given by $-\theta\Omega$, where θ is a constant. This corresponds to a delay in the time domain. Design methods for both types of filters are described in the next two sections.

IIR Filter Design

The two methods for designing IIR filters are termed *analog emulation* (or indirect design) and *direct design*. Analog emulation involves designing an analog filter first and then using one of the mapping techniques described in the section "s-Plane to z-Plane Mappings" to convert it to a digital filter. This method has advantages in that there is a wealth of design techniques for analog filters that can be used in digital filter design. Direct design methods generally involve numerical techniques, and they are often preferred over analog emulation when the sampling period is not very small. Direct design is beyond the scope of this handbook; consult reference [2] for more information on the topic.

Analog filter design begins by selecting a bandwidth, a prototype of filter, and an order of the filter. Additional specifications may be set on the amount of ripple that is allowed in the passband or stopband. Two common analog prototypes are the Butterworth filters and the Chebyshev filters.

Butterworth filter: The Butterworth filter is characterized by having no zeros and having poles that are situated on a semicircle in the left-half of the s-plane. The distance of the poles to the origin is the bandwidth frequency and is denoted as ω_b. The angle of the poles can be determined by equally spacing out twice the number of poles around a full circle with radius ω_b and then keeping only the poles in the left-half plane, as shown in Fig. 29.5. An Nth order Butterworth filter is given by

$$H(s) = \frac{\omega_b^N}{\prod_k (s - \omega_b p_k)}$$

where

$$p_k = \begin{cases} e^{jk\pi/N}, & k = \dfrac{N+1}{2} \text{ to } \dfrac{3N-1}{2} \text{ for } N \text{ odd} \\ e^{j(k+0.5)\pi/N}, & k = \dfrac{N}{2} \text{ to } \dfrac{3N-2}{2} \text{ for } N \text{ even} \end{cases}$$

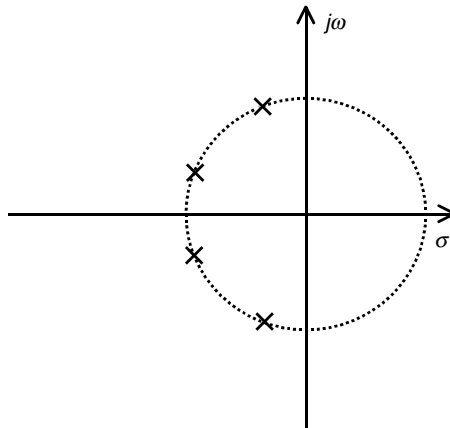



FIGURE 29.5 Pole distribution of a fourth-order Butterworth filter.

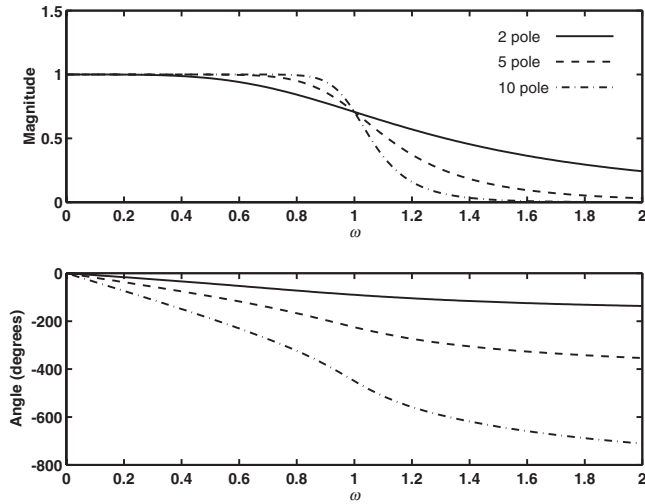



FIGURE 29.6 Comparison of analog Butterworth filters.

This filter is lowpass in that the magnitude of the frequency response is reasonably flat and close to a value of 1 for $\omega < \omega_b$ and drops off sharply beyond the bandwidth frequency. The larger the order of the filter, the sharper the drop off. Three filters are compared in Fig. 29.6. Notice that the sharp transitions offered by the larger order filters do come with a price: the phase is also decreased dramatically. The phasing becomes important in real-time measurement systems such as that required by feedback controllers.

Chebyshev filter: Unlike the monotonic behavior of the Butterworth filter, the Chebyshev filter allows some ripple in the magnitude plot for either the passband or the stopband. The Type 1 Chebyshev filter allows for ripple in the passband while the Type 2 Chebyshev filter allows for ripple in the stopband. Allowing for a ripple results in the Chebyshev filters having sharper transitions near the bandwidth than are achievable by a Butterworth filter of the same order. In Chebyshev design, the cutoff frequency ω_c is usually specified as opposed to the bandwidth. The cutoff frequency is the frequency at which the magnitude of the filter decays to a preset ratio of the DC value. When this ratio is 0.707, the cutoff frequency is the bandwidth. Often, in Chebyshev design, this ratio is chosen to correspond to the amount of ripple allowed in the passband. A Type 1 Chebyshev lowpass filter is defined by the relationships:

$$|H(\omega)| = \frac{1}{\sqrt{1+\varepsilon^2 C_N^2(\omega/\omega_c)}}$$

and

$$C_N(x) = 2xC_{N-1}(x) - C_{N-2}(x)$$

The $C_N(x)$ expression is called the Nth order Chebyshev polynomial, and it is calculated recursively starting with $C_0(x) = 1$ and $C_1(x) = x$. The value of $\varepsilon > 0$ determines the amount of ripple allowed in the passband; in particular, the ripple exists between the values of 1 and $1/\sqrt{1+\varepsilon^2}$. Consider, for example, the Type 1 Chebyshev filters shown in Fig. 29.7; these filters were designed to have 1 dB of ripple in the passband ($\varepsilon = 0.51$). Note that $\varepsilon = 1$ for 3 dB of ripple.

As mentioned above, these prototype filters are lowpass. To design another type of filter, first the lowpass filter is designed, $H(s)$, with a cutoff frequency ω_c (typically chosen to be 1). Then a frequency transformation is used to convert the filter to the desired type. The standard frequency transformations are given below.

Lowpass to lowpass: To obtain a lowpass filter with cutoff frequency ω_1, replace s in the original $H(s)$ by $s\omega_c/\omega_1$.

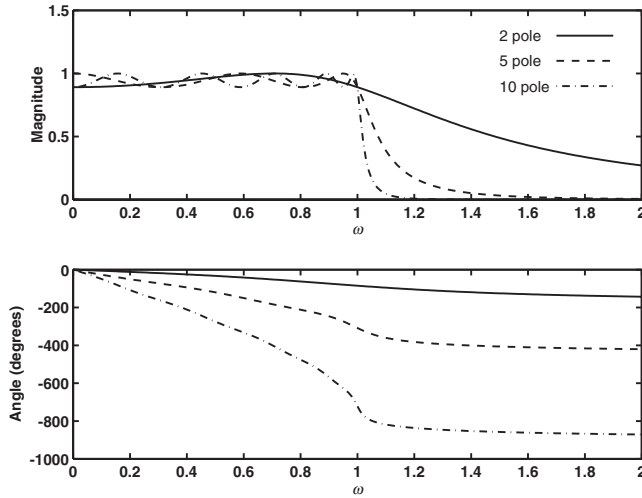



FIGURE 29.7 Comparison of analog Type I Chebyshev filters.

Lowpass to highpass: To obtain a highpass filter with a passband running from ω_1 to ∞, replace s in the original $H(s)$ by $\omega_1\omega_c/s$.

Lowpass to bandpass: To obtain a bandpass filter with a passband running from ω_1 to ω_2, replace s in the original $H(s)$ by

$$\frac{s^2 + \omega_2\omega_1}{s(\omega_2 - \omega_1)}.$$

Lowpass to bandstop: To obtain a bandstop filter with stopband running from ω_1 to ω_2, replace s in the original $H(s)$ by

$$\frac{s(\omega_2 - \omega_1)}{s^2 + \omega_2\omega_1}.$$

FIR Filter Design

One way to obtain an FIR filter is to truncate the impulse response of an ideal IIR filter. For example, an ideal IIR lowpass filter has the frequency response:

$$H(\Omega) = \begin{cases} A, & -\Omega_c \le \Omega \le \Omega_c \\ 0, & \text{otherwise} \end{cases}$$

where A is a constant and Ω_c is the cutoff frequency. The impulse response of this filter is found from taking the inverse DTFT using Eq. (29.3):

$$h[n] = \frac{A\Omega_c}{\pi} e^{jn\Omega_c/2} \operatorname{sinc}\left(\frac{\Omega_c n}{\pi}\right)$$

Notice that this has infinite duration for both $n < 0$ and $n > 0$. Creating an FIR filter would entail truncating the impulse response for $n < -N$ and for $n > N$. However, the original IIR filter and the resulting truncated FIR filter are both noncausal; that is, the impulse response is nonzero for $n < 0$.

Noncausal filters need future values of the input in order to calculate the present value of the output; hence, they cannot be implemented in real-time. For this reason, typical IIR design uses nonideal filters (often based on analog prototypes) that approximate the ideal frequency response. When filtering a signal off-line that has been stored, causality is no longer required since all of the values of the signal are available (including "future" values).

In order to perform real-time implementation of an FIR filter that was generated by truncating an ideal IIR filter, the filter must be delayed so that all of the significant information of the impulse response occurs for $n \geq 0$. This delay in the time domain is equivalent to a linear phase lag in the frequency domain.

Thus, an FIR filter can be designed by first selecting an ideal IIR filter (lowpass, highpass, etc.), then taking the inverse DTFT to find the impulse response, and then truncating the impulse response, and finally, delaying it in time. An equivalent and more preferred method is to rearrange the steps described above. First, add a phase lag in the frequency response of the ideal IIR filter. This is done by multiplying the frequency response by $e^{j(N-1)/2}$. Then, take the inverse DTFT and truncate it for $n < 0$ and $n > N - 1$. The result is a causal FIR filter with order N.

The following are generic FIR filters of order N that have been generated using the method described above. Let $m = (N - 1)/2$.

Lowpass FIR filter with cutoff frequency Ω_c:

$$h[n] = \begin{cases} \dfrac{\Omega_c}{\pi}, & n = 0 \\ \dfrac{\Omega_c}{\pi}\text{sinc}\left[\dfrac{\Omega_c(n-m)}{\pi}\right], & \text{for } 0 < n \leq N-1 \\ 0, & \text{otherwise} \end{cases}$$

where $\text{sinc}(x) = \sin(\pi x)/\pi x$.

Highpass FIR filter with passband from Ω_1:

$$h[n] = \begin{cases} 1 - \dfrac{\Omega_1}{\pi}, & \text{for } n = 0 \\ -\dfrac{\Omega_1}{\pi}\text{sinc}\left[\dfrac{\Omega_1(n-m)}{\pi}\right]0, & \text{for } 0 < n \leq N-1 \\ 0, & \text{otherwise} \end{cases}$$

Bandpass FIR filter with passband from Ω_1 *to* Ω_2:

$$h[n] = \begin{cases} \dfrac{\Omega_2 - \Omega_1}{\pi}, & \text{for } N = 0 \\ \dfrac{\Omega_2}{\pi}\text{sinc}[\Omega_2(n-m)/\pi] - \dfrac{\Omega_1}{\pi}\sin[\Omega_1(n-m)/\pi], & \text{for } 0 < n \leq N-1 \\ 0, & \text{otherwise} \end{cases}$$

To implement these filters, the coefficients in Eq. (29.5) are set to $b_m = h[m-1]$, $a_1 = 1$, and $a_m = 0$ for $m > 1$.

The FIR filters designed using this method have frequency responses that have rather sharp transitions between the passband and the stopband (the larger the order, the sharper the transition), but they tend to give rise to a ripple in the passband and stopband. This ripple results from the sharp truncation of the IIR filter's impulse response. A more gradual truncation using a window can be performed that smooths the ripple in the frequency response. The windows discussed in the section "*s*-Plane to *z*-Plane Mappings"

that are employed in data collection are also used in FIR filter design, where the modified filter is given as $h[n]w[n]$. FIR design using different windows is discussed in further detail in [1,2].

Computer-Aided Design of Digital Filters

Matlab™ is a common software package for signal processing analysis and design. The signal processing toolbox contains several commands for designing and simulating digital filters. For example, the commands butter and cheby1 automatically design a prototype analog filter for an IIR and then use the bilinear transformation to map the filter to the discrete-time domain. Lowpass, highpass, bandstop, and bandpass filters can be designed using these commands as long as the digital cutoff frequencies, normalized by π, are specified. To design a digital lowpass filter based on the analog Butterworth filter with cutoff frequency $w1$, use the command $[b, a]$ = butter(N, $w1*T$/pi) where N is the number of poles, T is the sampling period, and $w1*T$ is digital cutoff frequency. This command puts the coefficients of the filter, defined in Eq. (29.5), in vectors b and a in ascending order. To design a digital highpass filter with analog cutoff frequency $w1$, use the commands $[b, a]$ = butter(N, $w1*T$/pi, 'high'). To design a digital bandpass filter with analog passband from $w1$ to $w2$, define $w = [w1, w2]$ and use the command $[b, a]$ = butter (N, $w*T$/pi). To design a digital bandstop filter with stopband from $w1$ to $w2$, define $w = [w1, w2]$ and use the command $[b, a]$ = butter(N, $w*T$/pi, 'stop'). The design for an Nth order Type I Chebyshev filter is accomplished using the same methods as for butter except that "butter" is replaced by "cheby1."

The signal processing toolbox also provides commands for designing FIR filters. To obtain a lowpass FIR filter with length N and analog cutoff frequency $w1$, use the command h = fir1($N-1$, $w1*T$/pi). The resulting vector h contains the impulse response of the FIR where $h(1)$ is the value of $h[0]$. The values in the vector h also equal the coefficients of b in Eq. (29.5) in ascending order. (Recall, that $a_1 = 1$ and $a_m = 0$ for $m > 1$.) A length N highpass FIR filter with analog cutoff frequency $w1$ is designed by using the command h = fir1($N-1$, $w1*T$/pi, 'high'). A bandpass FIR filter with passband from $w1$ to $w2$ is obtained by typing h = fir1($N-1$, $w*T$/pi) where $w = [w1, w2]$. A bandstop FIR filter with stopband from $w1$ to $w2$ is obtained by typing h = fir1($N-1$, $w*T$/pi, 'stop') where $w = [w1, w2]$. The fir1 command uses the Hamming window by default. Other windows are obtained by adding an option of "hanning" or "boxcar" (which is the rectangular window) to the arguments; for example, h = fir1($N-1$, $w1*T$/ pi, 'high,' boxcar(N)) creates a highpass FIR filter with analog cutoff frequency $w1$ using a rectangular window.

The filter command in Matlab is used to compute an output of a digital filter given its input sequence. An example of its use is y = filter(b, a, x) where b and a are the coefficients of the filter and x is the input sequence.

Filtering Examples

Quite often, 60 Hz noise is encountered in measurements of electromechanical systems due to standard line voltage. (Note, in Europe noise at a 50-Hz frequency is typically encountered.) For demonstration purposes, a 60-Hz signal is superimposed on a lower frequency signal shown in Fig. 29.8. To alleviate the detrimental effects of the 60-Hz noise, a bandstop filter may be employed. Typically, most systems

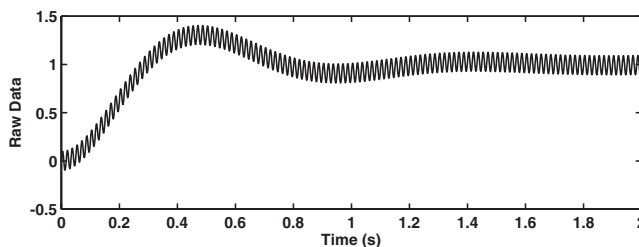



FIGURE 29.8 Measurement corrupted with 60-Hz noise.

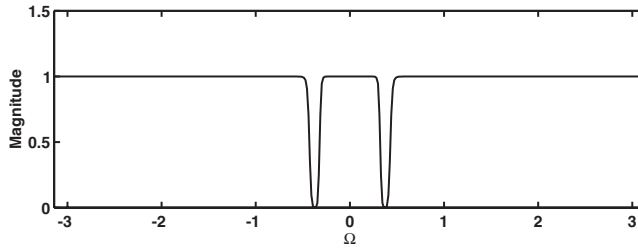



FIGURE 29.9 Bandstop filter.

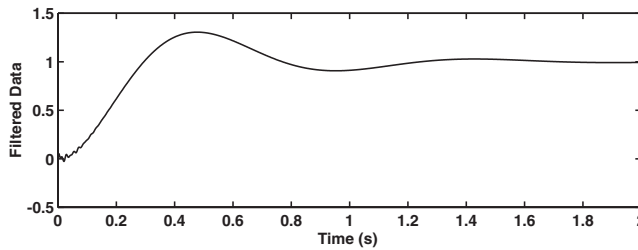



FIGURE 29.10 Filtered measurement.

have a bandstop filter designed around 60 Hz to avoid the type of response seen in Fig. 29.8. The following Matlab commands can be employed to design an eighth-order digital Butterworth bandstop filter whose break frequencies are 50 and 70 Hz. Thus, the filter should reject the 60-Hz noise.

```
T = 0.001; %Sample period
n = 4; % half the order of filter
low_freq = 50 * (2*pi); %Stop signals between 50 and 70 Hz
high_freq = 70 * (2*pi);

w1 = low_freq*(T/pi); % normalized digital break frequencies
w2 = high_freq*(T/pi);
w = [w1 w2];

[b,a] = butter(n,w,'stop'); % filter coefficients

W = -pi:pi/200:pi; % define a digital frequency vector
H = freqz(b,a,W); % computes the frequency response for plotting
```

Figure 29.9 shows the magnitude of the frequency response for the resulting IIR filter. Note that the frequency variable is plotted for the range $[-\pi, \pi]$ where DC frequency corresponds to $\Omega = 0$ and the highest frequency allowable is $\Omega = \pi$. In this example, the digital break frequencies correspond to $\Omega_1 = 50(2\pi)T = 0.314$ and $\Omega_2 = 70(2\pi)T = 0.44$. Figure 29.10 shows the result of applying this filter to the noisy signal. For all practical purposes, the 60-Hz noise is completely attenuated. As can be seen in Fig. 29.10, there are some initial system transients during the first 100 ms of the step response. This is a combination of the fourth-order Butterworth filter and the initial system transients to the 60-Hz signal. It should be noted that the sample frequency of 1000 kHz is fast enough to accurately capture the 60-Hz signal. If a sample frequency of less than 120 Hz is used, the 60-Hz signal will be aliased, and no amount of digital filtering would be able to eliminate the effects of the 60-Hz disturbance.

Another application of digital filtering in mechatronics is used when estimating displacement from an acceleration measurement. A simplistic approach to calculating the displacement is to integrate the acceleration twice. In the *s*-domain, this double integration is equivalent to multiplying by $1/s^2$. Using the

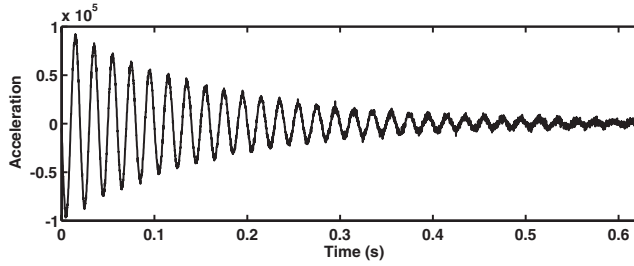



FIGURE 29.11 Acceleration measurement.

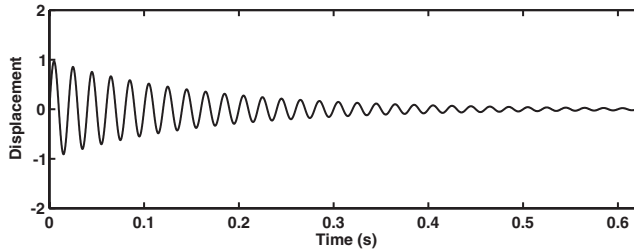



FIGURE 29.12 Actual displacement.

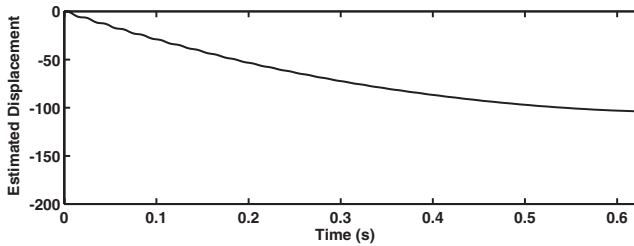



FIGURE 29.13 Estimated displacement without use of a prefilter.

bilinear transformation to convert $1/s^2$ to the z-domain yields the following transfer function,

$$H(z) = H(s)\Big|_{s=\frac{2(z-1)}{T(z+1)}} = \frac{1}{s^2}\Bigg|_{s=\frac{2(z-1)}{T(z+1)}} = \left(\frac{T^2}{4}\right)\left(\frac{z^2+2z+1}{z^2-2z+1}\right) = \left(\frac{T^2}{4}\right)\left(\frac{1+2z^{-1}+z^{-2}}{1-2z^{-1}+z^{-2}}\right)$$

The corresponding difference equation used to calculate the displacement $y[\bullet]$ from the acceleration $y_{dd}[\bullet]$ is $4(y[n] - 2y[n-1] + y[n-2]) = T^2(y_{dd}[n] + 2y_{dd}[n-1] + y_{dd}[n-2])$. However, accelerometers generally do not have good response at low frequencies; in fact, they often insert a bias in the data yielding a drift in the calculated displacement. They also are very sensitive to random vibrations. An alternate approach is to process the acceleration data through a bandpass filter before using the difference equation to integrate it numerically. The bandpass range must contain the natural frequencies in the system.

Consider, for example, the acceleration data shown in Fig. 29.11, which has some random noise. This signal is sampled at a rate of 6400 Hz, where the natural frequency of the system is 50 Hz. Figure 29.12 shows the actual displacement, while Fig. 29.13 shows the estimated displacement calculated by numerically integrating the acceleration data, using the difference equation given above. This estimation is very poor. Alternatively, an analog eighth-order Chebyshev Type I bandpass filter with passband 25–500 Hz

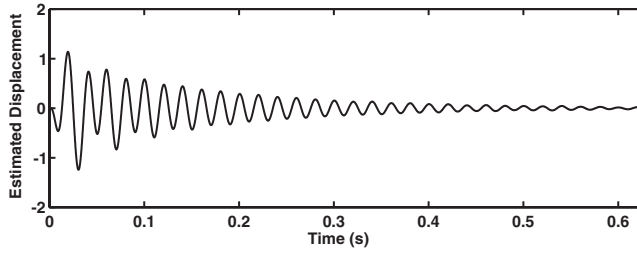



FIGURE 29.14 Estimated displacement with IIR prefilter.

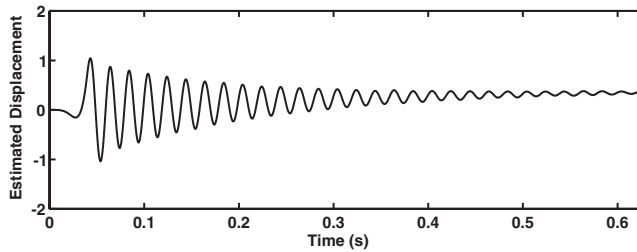



FIGURE 29.15 Estimated displacement with FIR prefilter.

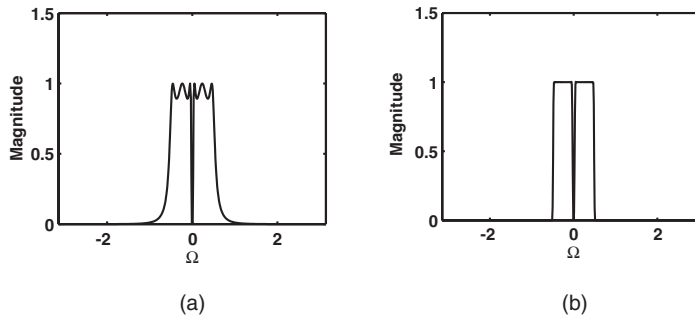



FIGURE 29.16 Digital bandpass filters: (a) Chebyshev IIR and (b) FIR filter.

is designed and then discretized using the bilinear transformation. The acceleration data is processed through this filter first, and then the filtered data are numerically integrated with the result shown in Fig. 29.14. Notice that the estimate is much better than that obtained without the bandpass filter. A 500th order FIR bandpass is also designed for this example with passband 25–500 Hz. After passing the data through the FIR filter, it is then numerically integrated resulting in the estimated displacement shown in Fig. 29.15. Due to the linear phase characteristic of the FIR filter, it has less transient distortion than the IIR filter, but it adds a larger lag. The larger the order, the more accurate the result, since less significant information is lost in truncating the impulse response of the ideal IIR bandpass filter, but the lag is larger. The magnitudes of the IIR filter and the FIR bandpass filters are shown in Fig. 29.16.

Two observations on this example should be mentioned:

1. The unfiltered calculation was extremely sensitive to bias in the data (as expected from the double integration). Therefore, the bias was removed from the acceleration before processing. Both filters effectively removed bias, so the results were virtually unchanged if the bias was present.
2. The FIR filter shows some drift. Presumably, the cause of the drift is that the filter seems to have some difficulty with the small stopband region near the origin. Increasing the stopband region

does reduce the drift. This can be done by decreasing the sample frequency or by increasing the passband frequency. Both of these remedies decrease the drift but increase other errors in the signal. Increasing the length of the filter decreases the drift error without introducing other errors.

Some of the Matlab commands used to design the filters and generate the results are:

```
[num,den] = c2dm(1,[1 0 0],T,'tustin'); %digitize 1/s^2
y1 = filter(num,den,ydd); % double integration of ydd

Wbreak = [2*pi*25*T, 2*pi*500*T}; % digital break frequencies
[b,a] = cheby1(4,1,Wbreak); % design IIR filter with 1dB ripple

W = -pi:pi/200:pi; % define digital frequency range for plot
H = freqz(b,a,W); % get frequency response
plot(W,abs(H)); % plot magnitude of frequency response

yddfilt = filter(b,a,ydd); % calculate output of IIR filter
y2 = filter(num,den,yddfilt); % double integration of yddfilt

hfir = fir1(500,Wbreak); % design FIR filter of order 500
yddfilt = filter(hfir,1,ydd); % calculate output of FIR filter
y3 = filter(num,den,yddfilt); % double integration of yddfilt
```

29.5 Digital Control Design

As in the digital filter design case, there are two general methods for designing a digital controller: an *indirect method* that is based on discretizing an analog design, and a *direct method* that is based on discretizing a plant (usually using the step-response matching method) and then designing the controller directly in the discrete domain. Most engineers learn classical continuous-time controls, and it is common for them to have more training in continuous-time control design than in discrete-time or digital control design. Fortunately, continuous-time control tools can often be used when designing digital control systems. To make use of controllers designed in the continuous-time domain, an s-plane to z-plane mapping is used. Any of the mappings discussed in this chapter can be used for a variety of controllers. It is always best to determine the mapping that is most efficient for a particular control or filter. Even though the bilinear approximation is more complex than the forward or backward approximations, it is used for most mechatronic systems. This is due to the fact that most modern controllers have enough computational power to manage the increased complexity at the required bandwidth of the mechatronic system.

As an example of the indirect design method, consider a PD (proportional derivative) controller that may be used to enhance the performance of a system. The derivative and proportional gains for the controller are K_d and K_p, respectively. The PD controller, $K(s)$, is given by

$$K(s) = K_d s + K_p. \tag{29.6}$$

Equation (29.6) can be implemented digitally using any of the s-plane to z-plane mappings discussed earlier in this chapter. As an example, the bilinear transformation is used generating the digital controller, $K(z)$.

$$K(z) = K(s)\big|_{s=\frac{2(z-1)}{T(z+1)}} = \frac{(2K_d + TK_p)z + (TK_p - 2K_d)}{Tz + T} \tag{29.7}$$

Besides the control gains, the only factor that is needed for Eq. (29.7) is T, the sample time. As previously stated, the sample time should be at least a factor of 5–10 times the fastest system time constant.

However, sampling times are often chosen to be several hundred times faster than the fastest system time constant. An alternative strategy for a feedback system is to choose the sample rate to be at least 20 times the desired closed loop bandwidth. Having sampling times that are substantially faster than the actual system mitigates any differences between the controller as it is designed in the continuous domain and the implementation in the discrete domain. It should be noted that as the sampling frequency becomes higher, the control gains become smaller. For example, in Eq. (29.7), as the sampling time becomes smaller, T becomes smaller requiring better numerical resolution for the controller gains. If T becomes smaller than the controller's numerical gain resolution, it may be erroneously implemented at a value of 0 (zero) yielding an incorrect control law.

Digital Control Example

Consider a high speed position motor with motor dynamics governed by the first-order equation

$$G(s) = \frac{\omega(s)}{V_{\text{in}}(s)} = \frac{K_m}{T_m s + 1}$$

where K_m is the motor gain constant, T_m is the motor time constant, $\omega(s)$ is the Laplace transform of the motor velocity, and $V_{\text{in}}(s)$ is the Laplace transform of the motor input voltage. To determine the values of T_m and K_m, the velocity step response of the motor is used. Figure 29.17 is the response of the motor to a 1-V step input. The motor gain, K_m, is the steady-state value of the final motor speed and is 5. This result can also be determined using the Final Value Theorem as

$$\begin{aligned}\lim_{t\to\infty}\omega(t) &= \lim_{s\to 0} s\omega(s) = \lim_{s\to 0} sG(s)V_{\text{in}}(s) \\ &= \lim_{s\to 0} sG(s)\frac{1}{s} = \lim_{s\to 0} s\frac{K_m}{T_m s + 1}\frac{1}{s} = K_m\end{aligned}$$

The motor time constant, T_m, can be computed by determining the motor velocity for the step response at time $t = T_m$ as follows:

$$\omega(t = T_m) = K_m(1 - e^{-t/T_m}) = K_m(1 - e^{-1}) = 0.632K_m$$

So the time required for the motor to reach 63.2% of its steady-state step response is its time constant. From Fig. 29.18, the time constant of this motor is 0.05 s. Thus, the transfer function for the motor is given by

$$G(s) = \frac{\omega(s)}{V_{\text{in}}(s)} = \frac{K_m}{T_m s + 1} = \frac{5}{0.05s + 1} \tag{29.8}$$

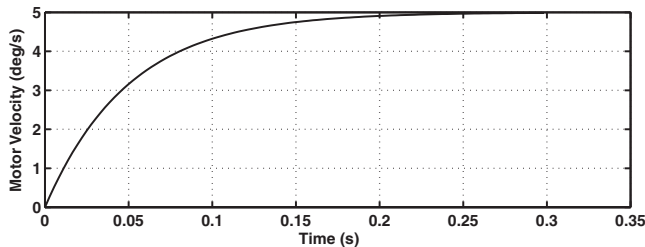



FIGURE 29.17 Motor velocity step response.

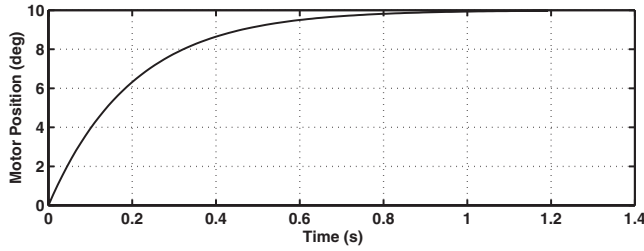



FIGURE 29.18 Closed-loop position response.

For this example, the motor is used in position control mode. Since the motor position is the integral of its velocity, Eq. (29.8) can be augmented with an integrator to generate the transfer function of the motor relating the input voltage to the output position, $\theta(s)$:

$$G_p(s) = \frac{\theta(s)}{V_{\text{in}}(s)} = \frac{K_m}{s(T_m s + 1)} \tag{29.9}$$

A PD controller is chosen for use in this example in order to enhance the system performance. To achieve a fast response with no overshoot, the derivative gain, K_d, and the proportional gain, K_p, are chosen to be 0.05 and 1, respectively, yielding the following control law:

$$K(s) = K_d s + K_p = 0.05s + 1 \tag{29.10}$$

Nominally, this design cancels the high frequency pole of the motor dynamics given in Eq. (29.9).

A sample period of 1 ms is chosen for this example as it is significantly faster than the system's time constants, and it is not an unreasonable value given modern digital controllers. As previously discussed, using a 1-kHz (1 ms) sample frequency mitigates any differences between the controller as it is designed in the continuous domain and its implementation is in the discrete domain. Using the bilinear transformation given in the section "*s*-Plane to *z*-Plane Mappings" results in a digital controller of the form:

$$K_D(z) = \frac{101z - 99}{z + 1}$$

In fact, the closed-loop response of the system using the digital controller cannot be easily distinguished from that of the system using the analog controller given by Eq. (29.10). The closed-loop position response of the motor for a 10° command input is shown in Fig. 29.18.

As mentioned in section the "Filtering Examples," 60 Hz noise is often present in measurements of electro-mechanical systems, so a bandstop filter is often used to attenuate the noise. In the closed-loop operation, the digital bandpass filter is cascaded with the digital PD controller.

References

1. Kamen, E.W., and Heck, B.S., *Signals and Systems Using the Web and Matlab,* 2nd ed., Prentice-Hall, Englewood Cliffs, NJ, 2000.
2. Britton Rorabaugh, C., *Digital Filter Designer's Handbook: with C++ Algorithms,* 2nd ed., McGraw-Hill, New York, 1997.

30

Control System Design Via $\mathcal{H}^2$ Optimization

Armando A. Rodriguez
Arizona State University

30.1 Introduction

This chapter addresses control system design via $\mathcal{H}^2$ (quadratic) optimization. A unifying framework based on the concept of a generalized plant and weighted optimization permits designers to address state feedback, state estimation, dynamic output feedback, and more general structures in a similar fashion. The framework permits one to easily incorporate design parameters and/or weighting functions that may be used to influence the outcome of the optimization, satisfy desired design specifications, and systematize the design process. Optimal solutions are obtained via well-known Riccati equations; e.g., Control Algebraic Riccati Equation (CARE) and Filter Algebraic Riccati Equation (FARE). While dynamic weighting functions increase the dimension of the Riccati equations being solved, solutions are readily obtained using today's computer-aided design software (e.g., MATLAB, robust control toolbox, μ-synthesis toolbox, etc.).

In short, $\mathcal{H}^2$ optimization generalizes all of the well-known quadratic control and filter design methodologies:

- Linear Quadratic Regulator (LQR) design methodology [7,11],
- Kalman–Bucy Filter (KBF) design methodology [5,6],
- Linear Quadratic Gaussian (LQG) design methodology [4,10,11].

$\mathcal{H}^2$ optimization may be used to systematically design constant gain state feedback control laws, state estimators, dynamic output controllers, and much more.

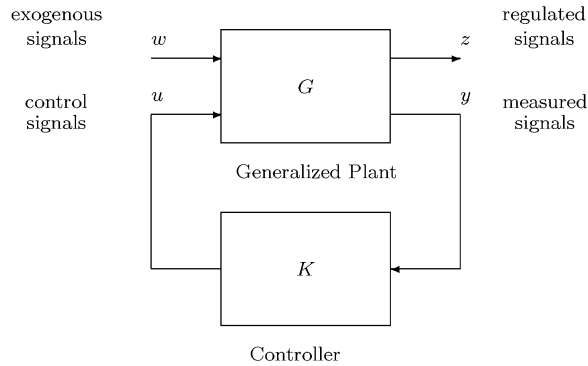



FIGURE 30.1 Generalized feedback system.

30.2 General Control System Design Framework

In this section, we present a general framework for control system (and estimator) design. Toward this end, we consider the generalized feedback system in Fig. 30.1. In this figure, G represents a *generalized plant*. G contains a model for the actual plant P (physical system) to be controlled. It may also contain additional (frequency dependent) weighting functions that are used to address closed loop design objectives. K represents a controller or compensator to be designed.

Central Idea: Design Via Optimization

The central idea here is that many important problems that arise in controls, estimation, filtering, and other areas of engineering may be cast in terms of a generalized plant G and a controller K to be designed by minimizing some norm (e.g., $\mathcal{H}^2$) on the closed loop transfer function matrix T_{wz} from the signals w to the signals z.

The Signals

To appreciate the flexibility of our generalized feedback system structure, it suffices to consider the nature of the signals z, u, w, and y in the figure. These signals may be described as follows:

- *Regulated Signals*. The signals $z \in \mathcal{R}^{n_z}$ represent *regulated signals* or signals that we would like to keep "small" in some sense, which depends on the application and desired performance objectives. Such signals might include tracking errors, actuator or control inputs, signal estimation errors, etc.
- *Control Signals*. The signals $u \in \mathcal{R}^{n_u}$ represent *control signals* or *manipulated variables* that are generated by the controller K. Control signals might include fuel flow to an engine, voltage applied to a dc motor, etc. They might also include state estimates provided by K. The idea is for K to manipulate and coordinate control signals u in a manner which keeps the regulated signals z "small." In practice, we typically have more signals that require "regulation" than controls (i.e., $n_z \geq n_u$). It should be noted, however, that generally if we want to independently control m quantities, then we need at least m independent controls. This basic tenet must be adhered to in practice. The more independent controls u that are available, the easier (in principle) it is to influence the signals z to be regulated.
- *Exogenous Signals*. The signals $w \in \mathcal{R}^{n_w}$ represent *exogenous* (*or external*) *signals* that act upon the system. Exogenous signals may include reference commands issued to the control system, disturbances acting on the system, sensor noise, etc.

- *Measurement Signals*. The signals $y \in \mathcal{R}^{n_y}$ represent measurements or signals that are directly available to the controller K. Measurements may include a portion of or all of the plant state variables, measurable plant "outputs," measurable control signals, measurable exogenous signals, etc. In practice, we typically have more exogenous signals than measurements (i.e., $n_w \geq n_y$). Generally, the more independent measurements we have the better—since, in theory, more useful information can be extracted.

Comment 30.1 (Toward a Separation Principle)

It is natural to associate the controls u with the regulated signals z. One might argue that the pair implicitly defines a regulation or control problem. This is analogous to the situation addressed in classical LQR problems. In such problems, one trades off control action (size) versus speed of regulation.

Similarly, it is natural to associate the exogenous signals w with the measurements y. One might argue that the pair implicitly defines an information extraction or estimation problem. This is analogous to the situation addressed in classical KBF problems. In such problems, one trades off sensor cost (or immunity to noise) versus speed of estimate construction.

Such associations suggest that just as in classical LQG problems, our surprisingly general structure may give rise to a natural separation principle. Indeed, this will be the case for the so-called $\mathcal{H}^2$ output feedback problem that we consider. ■

General $\mathcal{H}^2$ Optimization Problem

The so-called general $\mathcal{H}^2$ optimization problem may be stated as follows:

- Find a proper real-rational (finite dimensional) controller K that internally stabilizes G such that the $\mathcal{H}^2$ norm of the closed loop system transfer function matrix $T_{wz}(K)$ is minimized:

$$\min_{K} \|T_{wz}(K)\|_{\mathcal{H}^2} \tag{30.1}$$

where

$$\|F\|_{\mathcal{H}^2} \stackrel{\text{def}}{=} \sqrt{\frac{1}{2\pi}\int_{-\infty}^{\infty} trace\{F^H(j\omega)F(j\omega)\}d\omega} \tag{30.2}$$

$$= \sqrt{\int_0^{\infty} trace\{f^H(t)f(t)\}dt} \tag{30.3}$$

$$= \|f\|_{\mathcal{L}^2(\mathcal{R}_+)} \tag{30.4}$$

and f is the impulse response matrix associated with the transfer function matrix F.

Comment 30.2 (Use of Two Norm: Wide Band Exogenous Signals)

Noting that the two norm measures the energy of the response to an impulse and noting that the transform of unit Dirac delta function δ is unity, it follows that the two norm is appropriate when the exogenous signals w are wide band in nature. This can always be justified by introducing appropriate (low pass) filters within G. It should be noted that these ideas have stochastic interpretations as well. Instead of unit delta functions, one instead deals with white noise with unit intensity. ■

Comment 30.3 (Control and Estimation Problems)

Although we are seeking an $\mathcal{H}^2$ optimal controller, it must be noted that the generalized plant framework will enable the design of state estimators as well as dynamic and constant gain control laws. ■

Given the above problem statement, it is appropriate to recall the following elementary result:

Lemma 30.1 (Two Norm of a Stable System)

Consider a causal stable LTI strictly proper system $F = [A, B, C]$. It follows that

$$\|F\|_{\mathcal{H}^2} = \|f\|_{\mathcal{L}^2(\mathcal{R}^+)} = \sqrt{CL_cC^H} = \sqrt{B^H L_o B} \tag{30.5}$$

where L_c is the system controllability gramian and L_o is the system observability gramian. The controllability gramian

$$L_c \overset{\text{def}}{=} \int_0^\infty e^{At} BB^H e^{A^H t} dt \tag{30.6}$$

is the unique symmetric (at least) positive semi-definite solution of the algebraic Lyapunov equation

$$AL_c + L_c A^H + BB^H = 0 \tag{30.7}$$

L_c is positive definite if and only if (A, B) is controllable. The observability gramian

$$L_o \overset{\text{def}}{=} \int_0^\infty e^{A^H t} C^H C e^{At} dt \tag{30.8}$$

is the unique symmetric (at least) positive semi-definite solution of the algebraic Lyapunov equation

$$A^H L_o + L_o A + C^H C = 0 \tag{30.9}$$

L_o is positive definite if and only if (A, C) is observable. ■

Comment 30.4 ($\mathcal{H}^2$ Norm May Mislead—$\mathcal{L}^\infty$ Norm Is Important)

It is important to note that the $\mathcal{H}^2/\mathcal{L}^2$ norm (or energy) of a function may be very small, while the function itself may be very large in amplitude. Consider a tall thin pulse, for example. This observation is critical because there are many important cases in which we are very concerned with the height of a function—more so than its energy. A good example of this comes from classical Nyquist stability theory [2,8]. Nyquist taught us that the peak magnitude of the sensitivity function $S = 1/(1 + L)$ associated with a standard negative feedback loop is very important in terms of the feedback loop's stability robustness. A large sensitivity means that the Nyquist plot comes close to the critical −1 point—implying that a small perturbation (or unanticipated modeling error) may cause the closed loop system to go unstable. To assist us with this fundamental issue we may use frequency dependent weighting functions, but what we really need is a norm that directly addresses such concerns. This motivates the so-called $\mathcal{H}^\infty$ and $\mathcal{L}^\infty$ norms as well as $\mathcal{H}^\infty/\mathcal{L}^\infty$ control theory [4,11]. ■

Comment 30.5 (Computation of $\mathcal{H}^2$ Norm in MATLAB)

The $\mathcal{H}^2$ norm of a system $F = [A, B, C, D]$ may be computed using the following MATLAB command sequence:

```
lc = lyap (a, b*b')
twonorm = sqrt (trace(c*lc*c'))
```

or

```
lo = lyap (a', c*c')
twonorm = sqrt (trace(b'*lo*b)).
```

■

Generalized Plant

The generalized plant G is assumed to possess the following two-port state space structure:

$$G = \left[\begin{array}{c|c} G_{11} & G_{12} \\ \hline G_{21} & G_{22} \end{array}\right] = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & 0_{n_z \times n_w} & D_{12} \\ C_2 & D_{21} & 0_{n_y \times n_u} \end{array}\right] = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] \tag{30.10}$$

where $G_{ij}(s) = C_i(sI - A)^{-1}B_j$, $A \in \mathcal{R}^{n\times n}$, $B_1 \in \mathcal{R}^{n\times n_w}$, $B_2 \in \mathcal{R}^{n\times n_u}$, $C_1 \in \mathcal{R}^{n_z\times n}$, $C_2 \in \mathcal{R}^{n_y\times n}$, $D_{12} \in \mathcal{R}^{n_z\times n_u}$, $D_{21} \in \mathcal{R}^{n_u\times n_z}$.

Comment 30.6 (Weighting Functions: Satisfying Closed Loop Design Specifications)

As stated earlier, the generalized plant G may contain frequency dependent weighting functions as well as a model for the physical system P (plant) being controlled. Typically $P = G_{22} = [A, B_2, C_2]$. Weighting functions within G may be viewed as design parameters (mathematical knobs) that may be manipulated by a designer to influence the $\mathcal{H}^3$ problem in a manner which results in a controller that is not just optimal—a notion that is often irrelevant in practical applications—but which satisfies desired closed loop design specifications. Weighting functions may be used to weight (penalize) tracking errors, actuator and other signal levels, state estimation errors, etc. By making the weight on a signal large in a specific frequency range, we are indirectly telling the optimization problem to find a controller that makes the signal small in that frequency range. By making the weight on a signal small in a specific frequency range, we are indirectly conveying our willingness to tolerate a signal which is large in that frequency range. This idea can be illustrated via example. ■

Comment 30.7 (D_{11} = 0 Necessary, D_{22} = 0 Not Necessary)

$D_{11} = 0$ *Necessary.* Note that we have assumed that $D_{11} = 0$; i.e., there is no direct path from the exogenous signals w to the regulated signals z. This assumption is essential for the $\mathcal{H}^2$ norm of the closed loop transfer function T_{wz} to be finite. If $D_{11} \neq 0$, then $\|T_{wz}\|_{\mathcal{H}^2}$ will be infinite and the $\mathcal{H}^2$ problem will be ill-posed; i.e., make no sense. If we have a nonzero D_{11}, adding strictly proper filters on either w or z (e.g., $[1000/(s + 1000)]I$) will result in $D_{11} = 0$. In this sense, the assumption is not restrictive.

$D_{22} = 0$ *Not Necessary.* It has also been assumed that $D_{22} = 0$; i.e., the transfer function matrix D_{22} from controls u to measurements y is strictly proper. This assumption is very realistic since G_{22} (our plant P) is typically strictly proper in practice. If not, high frequency dynamics (e.g., actuator dynamics, flexible modes, parasitics, etc.) may be included to make it strictly proper. One might even include a simple high bandwidth low pass filter (e.g., $1000/(s + 1000)$) to make G_{22} strictly proper. If this is not desirable because of the increased dimension, there is an alternative that does not increase the dimension of G.

- One can always remove D_{22} from G_{22} to obtain a new generalized plant $\hat{G}$ with $\hat{D}_{22} = 0$. The term D_{22} is then absorbed into an augmented controller $\hat{K}$ by noting that u is related to y as follows:

$$u = K[y + D_{22}u] = [I - KD_{22}]^{-1}Ky \tag{30.11}$$

Noting this, it follows that the augmented controller, denoted $\hat{K}$, is given by

$$\hat{K} = [I - KD_{22}]^{-1}K \tag{30.12}$$

The $\mathcal{H}^2$ problem can then be carried out for $\hat{G}$ and $\hat{K}$ (without regard to D_{22}). When the optimal controller $\hat{K}$ for $\hat{G}$ is obtained, one can compute the optimal controller K for G using the relationship

$$K = \hat{K}[I + D_{22}\hat{K}]^{-1} \tag{30.13}$$

With this stated, the assumption $D_{22} = 0$ is made without any loss of generality. ■

Closed Loop Transfer Function Matrices

Given the structure for the generalized plant G, we have the following closed loop relationships:

$$u = Ky \tag{30.14}$$
$$= K(G_{21}w + G_{22}u) \tag{30.15}$$
$$= [I - KG_{22}]^{-1}KG_{21}w \tag{30.16}$$
$$= K[I - G_{22}K]^{-1}G_{21}w \tag{30.17}$$
$$y = [I - G_{22}K]^{-1}G_{21}w \tag{30.18}$$
$$z = G_{11}w + G_{12}u \tag{30.19}$$
$$= G_{11}w + G_{12}Ky \tag{30.20}$$
$$= [G_{11} + G_{12}K[I - G_{22}K]^{-1}G_{21}]w \tag{30.21}$$

From this, we have the following closed loop transfer function matrices:

$$T_{wu} = K[I - G_{22}K]^{-1}G_{21} \tag{30.22}$$
$$T_{wy} = [I - G_{22}K]^{-1}G_{21} \tag{30.23}$$
$$T_{wz} = G_{11} + G_{12}K[I - G_{22}K]^{-1}G_{21} \tag{30.24}$$

We say that each of these is a *linear fractional transformation* (*LFT*) involving K.

Comment 30.8 (Well Posedness of Closed Loop System)

In the above manipulation, it has been assumed that the inverse $[I - G_{22}K]^{-1}$ is well defined. This well posedness condition is guaranteed by our assumption that $D_{22} = 0$. This assumption implies that $G_{22}(j\infty) = D_{22} = 0$ and hence that the inverse is well defined. ■

The following example shows how to formulate a so-called Weighted $\mathcal{H}^3$ Mixed Sensitivity Problem to address feedback control system design issues.

Example 30.1 (Weighted $\mathcal{H}^2$ Mixed Sensitivity Problem: Design Philosophy)

This example considers the design of a controller K for a plant $P = [A_p, B_p, C_p, D_p]$ as shown in Fig. 30.2. To obtain K, we will formulate an $\mathcal{H}^2$ optimization that considers (directly or indirectly) various issues that are of importance in the design of a good feedback loop.

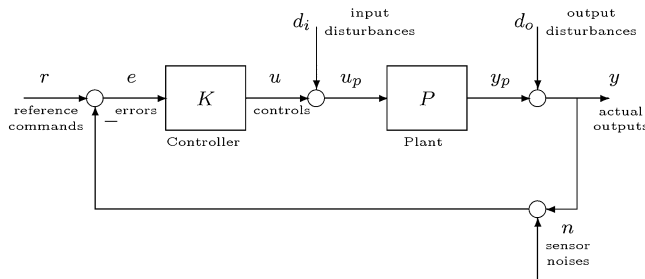



FIGURE 30.2 Standard negative feedback loop.

Feedback System Performance Issues. Generally, in designing a feedback controller K as shown in Fig. 30.2, a designer must consider each of the following closed loop performance issues:

- *Closed Loop Stability*. The closed loop system should be stable. This involves all closed loop transfer function matrices since we generally want all of them to be stable. A stricly proper closed loop transfer function matrix whose $\mathcal{H}^2$ norm is infinite, for example, implies that the transfer function matrix is unstable (or marginally stable). Stable strictly proper transfer function matrices necessarily have a finite $\mathcal{H}^3$ norm.
- *Command Following*. The closed loop system should exhibit good low frequency reference command following; i.e., the output y (not to be confused with generalized plant measurements) should track low frequency reference commands r that are issued to the feedback system. This typically requires that the sensitivity transfer function matrix

$$S \overset{\text{def}}{=} [I + PK]^{-1} \tag{30.25}$$

be small at low frequencies.
- *Disturbance Attenuation*. The closed loop system should exhibit good low frequency disturbance attenuation. For disturbances d_o modeled at the plant output, this requires that the sensitivity transfer function matrix be small at low frequencies. For disturbances d_i modeled at the plant input, this requires that

$$T_{d_i y} \overset{\text{def}}{=} SP \tag{30.26}$$

be small at low frequencies.
- *Sensor Noise Attenuation*. The closed loop system should exhibit good high frequency noise n attenuation. This typically requires that the complementary sensitivity transfer function matrix

$$T \overset{\text{def}}{=} I - S \tag{30.27}$$

be small at high frequencies.
- *Stability Robustness*. The closed loop system should exhibit robustness with respect to high frequency unmodeled dynamics (e.g., flexible modes, parasitic dynamics, time delays, etc.); This typically requires that the "peak" of some closed loop transfer function matrix be small at high frequencies.
  - Multiplicative Modeling Error. For a plant modeled as

$$P_{\text{act}} = [I + \Delta]P \tag{30.28}$$

  where P_{act} represents the actual plant, P represents a nominal model, and Δ represents a stable multiplicative perturbation at the plant output, the relevant closed loop transfer function matrix (that seen by Δ) is T.
  - Additive Modeling Error. For a plant modeled as

$$P_{\text{act}} = P + \Delta \tag{30.29}$$

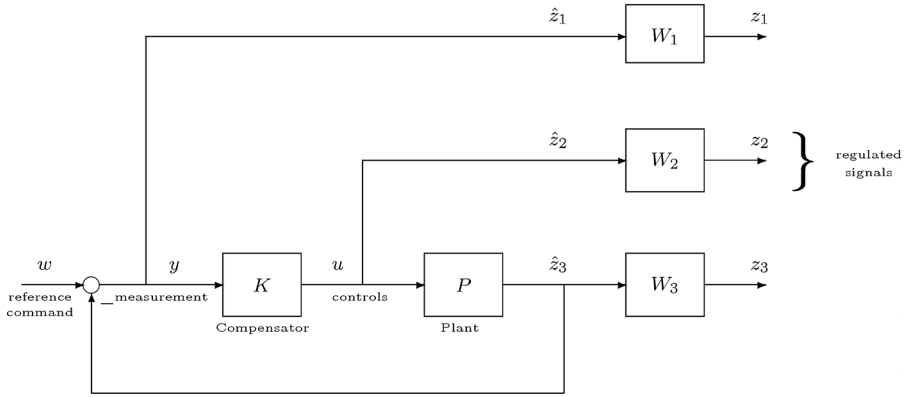



FIGURE 30.3 Negative feedback system for weighted mixed sensitivity problem.

where P_{act} represents the actual plant, P represents a nominal model, and Δ represents a stable additive perturbation, the relevant closed loop transfer function matrix (that seen by Δ) is KS.

- *Reasonable Control Action*. The closed loop system should exhibit reasonably sized control action for typical reference commands and sensor noise. This typically requires that the "size" of KS be controlled. Too much lead (i.e., derivative action) in K may help in terms of stabilization, achieving a high bandwidth and phase margin, but it may result in controls that are unnecessarily large in the presence of typical reference commands r and sensor noise n.

The above list suggests that there are many important issues that impact the control system design process. Part of a designer's job, however, is to prioritize and select issues that are most important. Toward this end, we turn our attention away from Fig. 30.2 and consider instead the "fictitious" (mathematical) system depicted in Fig. 30.3.

Weighting Functions and Closed Loop Transfer Function Matrix

Figure 30.3 includes specific weighting functions that will help us formulate an $\mathcal{H}^2$ optimization that (directly or indirectly) addresses some of the issues mentioned above. The figure shows a weighting W_1 on the signal y (the tracking error), a weighting W_2 on the controls u, and a weighting W_3 on the plant outputs $\hat{z}_3$. From the figure, it follows that regulated signals

$$z = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

are related to the exogenous signals w as follows:

$$z_1 = W_1 \hat{z}_1 = W_1 S w \tag{30.30}$$

$$z_2 = W_2 \hat{z}_2 = W_2 K S w \tag{30.31}$$

$$z_3 = W_3 \hat{z}_3 = W_3 T w \tag{30.32}$$

From this, it follows that the closed loop transfer function matrix from w to z is given by

$$T_{wz} = \begin{bmatrix} W_1 S \\ W_2 KS \\ W_3 T \end{bmatrix} \tag{30.33}$$

Since T_{wz} involves various "sensitivity" transfer function matrices, we say that we have a weighted mixed sensitivity problem.

Selection of Weighting Functions

Typically, the weighting functions W_1, W_2, W_3 are selected to be stable transfer function matrices that are (at least initially) diagonally structured.

- *Sensitivity Weighting*. One might select the sensitivity weighting W_1 on the sensitivity S as follows:

$$W_1 = \left[\frac{k_1}{s+\varepsilon}\right] I_{n_y \times n_y} \tag{30.34}$$

  where k_1, $\epsilon > 0$. The parameter k_1 is typically selected to be large. The parameter ϵ is typically selected to be small. Such selections are made so that S is heavily penalized at low frequencies—precisely where we want K to make S small.

- *Control Weighting*. One might select the control weighting W_2 on KS as follows:

$$W_2 = k_2 I_{n_u \times n_u} \tag{30.35}$$

  where $k_2 > 0$ provides a nonsingular penalty on the controls u (i.e., on KS).

- *Output Weighting*. One might select the output weighting W_3 on T as follows:

$$W_3 = \left[\frac{k_3(s+z_3)}{s+p_3}\right] I_{n_y \times n_y} \tag{30.36}$$

  with $k_3 > 0$ and $z_3 < p_3$. Such a weighting would penalize T more heavily at higher frequencies.

In general, care must be taken in selecting the structure of weighting functions. Inappropriate selections may result in an ill-posed problem and a very arduous design process. For example, W_1 must be strictly proper for $\mathcal{H}^2$ problems—otherwise the $\mathcal{H}^2$ norm of $W_1 S$ makes no sense (since S approaches the identity at high frequencies). While there exists no precise systematic method for the selection of weighting functions, the above structures seem to work well (as starting points) in many applications.

Input–Output Representation for Generalized Plant G

To obtain an input–output (transfer function matrix) description for our generalized plant, we must express the regulated signals z_1, z_2, z_3 and the measurements y in terms of the exogenous signals w and the controls u. Doing so yields

$$z_1 = W_1 \hat{z}_1 = W_1(w - Pu) = W_1 w - W_1 P u \tag{30.37}$$

$$z_2 = W_2 \hat{z}_2 = W_2 u \tag{30.38}$$

$$z_3 = W_3 \hat{z}_3 = W_3 P u \tag{30.39}$$

$$y = w - \hat{z}_3 = w - Pu \tag{30.40}$$

From this, we obtain the following input–output (transfer function matrix) description for our generalized plant G:

$$\left[\frac{z}{y}\right] = \left[\begin{array}{c|c} G_{11} & G_{12} \\ \hline G_{21} & G_{22} \end{array}\right] \left[\frac{w}{u}\right] \tag{30.41}$$

or

$$\begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \hline y \end{bmatrix} = \left[\begin{array}{c|c} W_1 & -W_1 P \\ 0 & W_2 \\ 0 & W_3 P \\ \hline I & -P \end{array}\right] \begin{bmatrix} w \\ u \end{bmatrix} \tag{30.42}$$

State Space Representation for Generalized Plant G

Next we obtain a two-port state space representation for G. To do so, we assume the following state space representations:

$$P = [A_p, B_p, C_p, D_p] \quad \text{with state } x_p \tag{30.43}$$

$$W_1 = [A_1, B_1, C_1, D_1] \quad \text{with state } x_1 \tag{30.44}$$

$$W_2 = [A_2, B_2, C_2, D_2] \quad \text{with state } x_2 \tag{30.45}$$

$$W_3 = [A_3, B_3, C_3, D_3] \quad \text{with state } x_3 \tag{30.46}$$

To obtain the desired state space representation for G, we need to express the signals $(\{\dot{x}_i\}_{i=1}^3, \dot{x}_p, \{z_i\}_{i=1}^3, y)$ in terms of the signals $(\{x_i\}_{i=1}^3, x_p, w, u)$. This is just a matter of simple bookkeeping. Doing so yields the following:

$$\dot{x}_1 = A_1 x_1 + B_1 y = A_1 x_1 + B_1(w - C_p x_p - D_p u) = A_1 x_1 - B_1 C_p x_p - B_1 D_p u \tag{30.47}$$

$$\dot{x}_2 = A_2 x_2 + B_2 u \tag{30.48}$$

$$\dot{x}_3 = A_3 x_3 + B_3 \hat{z}_3 = A_3 x_3 + B_3(C_p x_p + D_p u) = A_3 x_3 + B_3 C_p x_p + B_3 D_p u \tag{30.49}$$

$$\dot{x}_p = A_p x_p + B_p u \tag{30.50}$$

$$z_1 = C_1 x_1 + D_1 y = C_1 x_1 + D_1(w - C_p x_p - D_p u) = C_1 x_1 - D_1 C_p x_p + D_1 w - D_1 D_p u \tag{30.51}$$

$$z_2 = C_2 x_2 + D_2 u \tag{30.52}$$

$$z_3 = C_3 x_3 + D_3 \hat{z}_3 = C_3 x_3 + D_3(C_p x_p + D_p u) = C_3 x_1 + D_3 C_p x_p + D_3 D_p u \tag{30.53}$$

$$y = w - C_p x_p - D_p u = -C_p x_p + w - D_p u \tag{30.54}$$

The above equations may be written in standard two-port form:

$$\begin{bmatrix} \dot{x} \\ \hline z \\ \hline y \end{bmatrix} = \left[\begin{array}{c|c|c} A & B_{11} & B_{12} \\ \hline C_{11} & D_{11} & D_{12} \\ \hline C_{21} & D_{21} & D_{22} \end{array}\right] \begin{bmatrix} x \\ \hline w \\ \hline u \end{bmatrix} \tag{30.55}$$

as follows:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_p \\ \hline z_1 \\ z_2 \\ z_3 \\ \hline y \end{bmatrix} = \left[\begin{array}{cccc|c|c} A_1 & & & -B_1C_p & & -B_1D_p \\ & A_2 & & & & B_2 \\ & & A_3 & B_3C_p & & B_3D_p \\ & & & A_p & & B_p \\ \hline C_1 & & & -D_1C_p & D_1 & -D_1D_p \\ & C_2 & & & & D_2 \\ & & C_3 & D_3C_p & & D_3D_p \\ \hline & & & -C_p & I & -D_p \end{array}\right] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_p \\ \hline w \\ \hline u \end{bmatrix} \tag{30.56}$$

Checking Assumptions

In selecting the weights, W_1, W_2, W_3, one must make sure that none of the "standard" $\mathcal{H}^2$ problem assumptions are violated. Thus far, we require that $D_{11} = 0$ and $D_{22} = 0$. To ensure that $D_{11} = 0$, we need

$$D_1 = 0 \tag{30.57}$$

To ensure that $D_{22} = 0$, we need

$$D_p = 0. \tag{30.58}$$

This results in

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_p \\ \hline z_1 \\ z_2 \\ z_3 \\ \hline y \end{bmatrix} = \left[\begin{array}{cccc|c|c} A_1 & & & -B_1C_p & & B_2 \\ & A_2 & & & & \\ & & A_3 & B_3C_p & & \\ & & & A_p & & B_p \\ \hline C_1 & & & & & \\ & C_2 & & & & D_2 \\ & & C_3 & \Delta_3C_p & & \\ \hline & & & -C_p & I & \end{array}\right] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_p \\ \hline w \\ \hline u \end{bmatrix} \tag{30.59}$$

In subsequent sections, additional assumptions will be imposed on the two-port state space representation for the generalized plant G. The additional assumptions imposed will depend upon the specific $\mathcal{H}^2$ problem being considered.

Weighted $\mathcal{H}^2$ Optimal Mixed Sensitivity Problem

Given the above, the weighted $\mathcal{H}^2$ optimal mixed sensitivity problem is then to find a real-rational (finite-dimensional) proper internally stabilizing controller K that minimizes $\|T_{wz}\|_{\mathcal{H}^2}$; i.e.,

$$\min_K \|T_{wz}\|_{\mathcal{H}^2} = \min_K \left\| \begin{bmatrix} W_1S \\ W_2KS \\ W_3T \end{bmatrix} \right\|_{\mathcal{H}^2} \tag{30.60}$$

We will show how this optimal control problem—and problems like it—can be readily solved using computer-aided design soflware (e.g., MATLAB, robust control toolbox, μ-tools). ■

Comment 30.9 (Construction of Generalized Plant G)

Generalized plants G are very easy to construct within SIMULINK. Input port blocks may be used to specify exogenous signals w and controls u. Output port blocks may be used to specify regulated signals z and measurements y. The "linmod" command may be applied to the constructed block diagram (SIMULINK file) to obtain a two-port state space (A, $B = [B_1\ B_2]$, $C = [C_1; C_2]$, $D = [D_{11}\ D_{12}; D_{21}\ D_{22}]$) representation for G. The syntax for the command is as follows:

```
[ a, b, c, d ] = linmod ('filename')
```

This method enables one to create generalized plant models quickly. ■

Overview of $\mathcal{H}^2$ Optimization Problems to Be Considered

Three fundamental problems are considered in this chapter:

1. **$\mathcal{H}^2$ Output Feedback Problem.** The solution to this problem is an optimal model-based dynamic compensator possessing the structure

$$K_{\text{opt}} = \left[\begin{array}{c|c} A - B_2 G_c - H_f C_2 & H_f \\ \hline -G_c & O_{n_u \times n} \end{array}\right] \tag{30.61}$$

   where G_c, is a control gain (state feedback) matrix and H_f is a filter gain (observer) matrix. G_c is found by using the solution of a Control Algebraic Riccati Equation (CARE)—similar to that found in Linear Quadratic Regulator (LQR) problems. H_f is found by using the solution of a Filter Algebraic Riccati Equation (FARE)—similar to that found in Kalman–Bucy Filtering (KBF) problems. The structure of K_{opt}, G_c, and H_f can be thought of as the solution to a classical Linear Quadratic Gaussian (LQG) control problem which gives rise to the well known separation principle: closed loop poles are the eigenvalues of $A - B_2G_c$, and the eigenvalues of $A - H_fC_2$.
2. **$\mathcal{H}^2$ State Feedback Problem**. The solution to this problem is an optimal constant gain (state feedback) compensator possessing the structure

$$K_{\text{opt}} = -G_c \tag{30.62}$$

   where G_c is a control gain (state feedback) matrix found by using the solution to a CARE—similar to that found in LQR problems. The poles of the resulting closed loop system are the eigenvalues of $A - B_2G_c$. In short, this problem should be viewed as a mechanism for computing control gain matrices G_c that may be used in a state feedback application or in a model-based compensator application.
3. **$\mathcal{H}^2$ Output Injection Problem.** The solution to this problem is an optimal constant gain (static) compensator possessing the structure

$$K_{\text{opt}} = -H_f \tag{30.63}$$

   where H_f is a filter gain (observer) matrix found by using the solution to a FARE—similar to that found in KBF problems. The poles of the resulting closed loop system are the eigenvalues of $A - H_fC_2$. In short, this problem should be viewed as a mechanism for computing filter gain matrices H_f that may be used in a state estimation application or in a model-based compensator application.

30.3 $\mathcal{H}^2$ Output Feedback Problem

In this section, we consider the $\mathcal{H}^2$ output feedback problem. This problem results in model-based (dynamic) compensators involving a control gain (state feedback) matrix G_c and a filter gain (observer) matrix H_f. As such, the problem generalizes the ideas presented in classical LQG theory.

The following "standard" $\mathcal{H}^2$ output feedback problem assumption is now made.

Assumption 30.1 ($\mathcal{H}^2$ Output Feedback Problem)

Throughout this section, it will be assumed that

1. *Plant G_{22} Assumption.* (A, B_2, C_2) stabilizable and detectable.
   - This assumption is necessary and suficient for the existance of a proper internally stabilizing controller K. With this assumption, the following model based (observer based) controller stabilizes the feedback loop in Fig. 30.1:

$$K = \left[\begin{array}{c|c} A - B_2 G_c - H_f(C_2 - D_{22} G_c) & H_f \\ \hline -G_c & O_{n_u \times n} \end{array}\right] \tag{30.64}$$

   provided that $(A - B_2G_c)$ and $(A - H_fC_2)$ are stable, as suggested by the classical separation principle from the theory of linear systems.

   This assumption mandates that *all of the "bad" open loop poles* (*right half plane and imaginary*) *must be controllable through the controls u and observable through the measurements y.*

   Suppose that G satisfies the assumption. Consider the augmentation of an integrator I/s (weighting function). Such an augmentation can result in the assumption being violated. Absorbing an integrator I/s (weighting function) on the exogenous signals w into G, for example, would violate the stabilizability assumption since it would introduce an open loop pole on the imaginary axis that is not controllable through the controls u. Absorbing an integrator on the regulated signals z into G would violate the detectability assumption since it would introduce an open loop pole on the imaginary axis that is not observable through the measurements y. Using $I/(s + \epsilon)(\epsilon > 0)$ instead of I/s—in either case—would result in a G that does not violate the assumption.
2. *Nonsingular Control Weighting Assumption.* $R = D_{12}^T D_{12} > 0$.
   - This assumption implies that $D_{12} \in \mathcal{R}^{n_z \times n_u}$ has full column rank (i.e., rank $D_{12} = n_u$) and hence that every control (direction) u influences the regulated signals z through D_{12} (i.e., D_{12} has no right null space). The matrix D_{12} must therefore be "tall" and "thin;" i.e.,

$$\text{(number of regulated signals) } n_z \geq n_u \text{ (number of control signals)} \tag{30.65}$$

   The matrix $R = D_{21}^T D_{12} \in \mathcal{R}^{n_u \times n_u}$ may be interpreted as a weighting on the controls u—just like the control weighting matrix "R" in LQR problems. As in the LQR problem, we might say that the control weighting R on u is nonsingular. The larger R, the smaller we want our controls to be—sacrificing speed of regulation. A large R results in a low "regulation" bandwidth. The smaller R, the larger we will permit our controls u to be—in order to speed up regulation. A small R results in a high "regulation" bandwidth.
3. *Regulator Assumption.* $\begin{bmatrix} j\omega I - A & -B_2 \\ C_1 & D_{12} \end{bmatrix}$ has full column rank $(n + n_u)$ for all ω.
   - This assumption implies that transfer function matrix from control signals u to regulated signals z has no (right) zero on the imaginary axis. Together with (1) and (2), it will guarantee that the Hamiltonian $\mathcal{H}_{\text{con}}$ involving (A, B_2, C_1, D_{12}, R)—that is, associated with the controls u and regulated signals z—will belong to *dom*(*Ric*). This, in turn, guarantees that the solution of the associated CARE results in a control gain matrix $G_c \in \mathcal{R}^{n_u \times n}$ such that $A - B_2G_c$ is stable.

The assumption implies that G has no imaginary modes that are unobservable through the regulated signals z; that is, all open loop poles on the imaginary axis must be observable through the regulated signals z. (A,C_1), therefore, cannot possess unobservable imaginary modes. This is a necessary condition. It is not suficient. An integrator hanging on the measurements y, for example, would violate this.

Since D_{12} has full column rank, the assumption is equivalent to the pair

$$(A - B_2R^{-1}D_{12}^TC_1,\ (I - D_{12}R^{-1}D_{12}^T)C_1) \tag{30.66}$$

having no unobservable imaginary modes.

- If D_{12} is square, then it is invertible and the assumption is equivalent to $A - B_2R^{-1}D_{12}^TC_1$ having no imaginary modes.
- If $D_{12}^TC_1 = 0$ (no cross penalty between controls and states), then the assumption is equivalent to (A, C_1) having no unobservable imaginary modes.

4. Nonsingular Measurement Weighting Assumption. $\Theta = D_{21}D_{12}^T > 0$.
   - This assumption implies that $D_{21} \in \mathcal{R}^{n_y \times n_w}$ has full row rank (i.e., rank $D_{21} = n_y$) and hence that the measurements y are linearly independent through D_{21} (i.e., D_{21} has no left null space). The matrix D_{21} must therefore be "short" and "fat;" i.e.,

$$\text{(number of measurements) } n_y \leq n \text{ (number of exogenous signals)} \tag{30.67}$$

   The matrix $\Theta = D_{21}D_{21}^T \in \mathcal{R}^{n_y \times n_y}$ may be interpreted as the intensity of sensor noise impacting the measurements y—just like the sensor noise intensity matrix "Θ" found in KBF problems. As in the KBF problem, we say that the intensity matrix Θ associated with the measurements y is nonsingular. The larger Θ, the more we want to low pass filter the measurements y—sacrificing speed of estimation. A large Θ results in a low bandwidth for the associated estimator (observer). The smaller Θ, the less we want to low pass filter the measurements y—trading off our immunity to noise for speed of estimation. A small Θ results in a high bandwidth for the associated estimator (observer).

5. *Filter Assumption.* $\begin{bmatrix} j\omega I - A & -B_1 \\ C_2 & D_{21} \end{bmatrix}$ has full row rank $(n + n_y)$ for all ω.
   - This assumption implies that transfer function matrix from exogenous signals w to measurements y has no (left) zero on the imaginary axis. Together with (1) and (3), it will guarantee that the Hamiltonian $\mathcal{H}_{\text{fil}}$ involving $(A, B_1, C_2, D_{21}, \Theta)$—that is, involving exogenous signals z and measurements y—will belong to *dom*(*Ric*). This, in turn, guarantees that the solution of the associated FARE results in a filter gain matrix $H_f \in \mathcal{R}^{n \times n_y}$ such that $A - H_fC_2$ is stable.

   The assumption implies that G has no imaginary modes that are uncontrollable through the exogenous signals w; that is, all open loop poles on the imaginary axis must be controllable through the exogenous signals w. (A, B_1), therefore, cannot possess uncontrollable imaginary modes. This is a necessary condition. It is not suficient. An integrator hanging on the controls u, for example, would violate this.

   Since D_{21} has full row rank, the assumption is equivalent to the pair

$$(A - B_1D_{21}^T\Theta^{-1}C_2,\ B_1(I - D_{21}^T\Theta^{-1}D_{21})) \tag{30.68}$$

having no uncontrollable imaginary modes.

- If D_{21} is square, then it is invertible and the assumption is equivalent to $A - B_1D_{21}^T\Theta^{-1}C_2$ having no imaginary modes.
- If $B_1D_{21}^T = 0$ (uncorrelated process and sensor noise), then the assumption is equivalent to (A, B_1) having no uncontrollable imaginary modes. ■

Comment 30.10 (Duality Relationships)

In the above discussion, we note the following dual relationships:

$$A \longleftrightarrow A^T \tag{30.69}$$

$$B_2 \longleftrightarrow C_2^T \tag{30.70}$$

$$C_1 \longleftrightarrow B_1^T \tag{30.71}$$

$$D_{12} \longleftrightarrow D_{21}^T \tag{30.72}$$

$$R = D_{12}^T D_{12} \longleftrightarrow \Theta = D_{21} D_{21}^T \tag{30.73}$$

These imply that

- controls u are dual to measurements y
- Regulated signals z are dual to exogenous signals w. ■

Hamiltonian Matrices

Associated with our $\mathcal{H}^2$ optimal control problem are the following two Hamiltonian matrices:

$$H_{\text{con}} = \begin{bmatrix} A & 0 \\ -C_1^T C_1 & -A^T \end{bmatrix} - \begin{bmatrix} B_2 \\ -C_1^T D_{12}^T \end{bmatrix} R^{-1} [D_{12}^T C_1 \quad B_2^T] \tag{30.74}$$

$$= \begin{bmatrix} A - B_2 R^{-1} D_{12}^T C_1 & -B_2 R^{-1} B_2^T \\ -C_1^T (I - D_{12}^T R^{-1} D_{12}^T) C_1 & -(A - B_2 R^{-1} D_{12}^T C_1)^T \end{bmatrix} \tag{30.75}$$

$$H_{\text{fil}} = \begin{bmatrix} A^T & 0 \\ -B_1 B_1^T & -A \end{bmatrix} - \begin{bmatrix} C_2^T \\ -B_1 D_{21}^T \end{bmatrix} \Theta^{-1} \begin{bmatrix} D_{21} B_1^T & C_2 \end{bmatrix} \tag{30.76}$$

$$= \begin{bmatrix} (A - B_1 D_{21}^T \Theta^{-1} C_2)^T & -C_2^T \Theta^{-1} C_2 \\ -B_1 (I - D_{21}^T \Theta^{-1} D_{21}) B_1^T & -(A - B_1 D_{21}^T \Theta^{-1} C_2) \end{bmatrix} \tag{30.77}$$

The first Hamiltonian is associated with an optimal state feedback control or regulator problem. The second is associated with an optimal filtering or estimation problem.

The solution to the $\mathcal{H}^2$ output feedback problem is now given [11, pp. 261–262].

Theorem 30.1 (Solution to $\mathcal{H}^2$ Output Feedback Problem Subject to Standard Assumptions)

Suppose that G satisfies the assumptions given in Assumption 30.1—the so-called standard $\mathcal{H}^2$ output feedback problem assumptions. Given this, we have the following.

The unique minimizing $\mathcal{H}^2$ optimal controller is n dimensional (like generalized plant G) and is given by

$$K_{\text{opt}} = \left[\begin{array}{c|c} A - B_2 G_c - H_f C_2 & H_f \\ \hline -G_c & O_{n_u} \times n_y \end{array}\right] \tag{30.78}$$

where the control gain matrix $G_c \in \mathcal{R}^{n_u \times n}$ *is given by*

$$G_c = R^{-1}[B_2^T X + D_{12}^T C_1] \tag{30.79}$$

$X = Ric(\mathcal{H}_{con}) \geq 0$ *is the unique (at least) positive semi-definite solution of the CARE:*

$$(A - B_2R^{-1}D_{12}^TC_1)^TX + X(A - B_2R^{-1}D_{12}^TC_1) + C_1^T(1 - D_{12}^TR^{-1}D_{12}^T)C_1 - XB_2R^{-1}B_2^TX = 0 \tag{30.80}$$

and the filter gain matrix $H_f \in \mathcal{R}^{n \times n_y}$ *is given by*

$$H_f = [YC_2^T + B_1D_{21}^T]\Theta^{-1} \tag{30.81}$$

$Y = Ric(\mathcal{H}_{fil}) \geq 0$ *is the unique (at least) positive semi-definite solution of the FARE:*

$$(A - B_1D_{12}^T\Theta^{-1}C_2)Y + Y(A - B_1D_{21}^T\Theta^{-1}C_2)^T + B_1(I - D_{21}^T\Theta^{-1}D_{21})B_1^T - YC_2^T\Theta^{-1}C_2Y = 0 \tag{30.82}$$

Moreover, the minimum norm is given by

$$\|T_{wz}(K_{\text{opt}})\|_{\mathcal{L}^2} = \sqrt{\|M_cB_1\|_{\mathcal{L}^2}^2 + \|R^{1/2}G_cM_f\|_{\mathcal{L}^2}^2} \tag{30.83}$$

$$= \sqrt{trace(B_1^TXB_1) + trace(RG_cYG_c^T)} \tag{30.84}$$

where

$$M_c = [A - B_2G_c, I_{n\times n}, C_1 - D_{12}G_c] \tag{30.85}$$

$$M_f = [A - H_fC_2, I_{n\times n}, B_1 - H_fD_{21}] \tag{30.86}$$

Finally, the closed loop poles are the eigenvalues of $A - B_2G_c$ *and* $A - H_fC_2$*.* ■

Comment 30.11 (Computing Optimal $\mathcal{H}^2$ Controller in MATLAB)

The following MATLAB command sequence may be used to compute the optimal $\mathcal{H}^2$ controller K_{opt} and the resulting closed loop transfer function matrix T_{wz}:

```
tss_g = mksys(a, [b1 b2], [c1; c2], [0*ones(nz, nw) d12; d21 0*ones(ny,
nu),'tss')
[ss_k ss_twz] = h21qg(tss_g, 'schur')
[a_k, b_k, d_k] = branch(ss_k, 'a,b,c,d')
```

The "mksys" command packs the two-port state space data for the generalized plant G into a column vector data structure (called a tree) possessing the "tss" (two-port state space) variable designation. All dimension information is encoded into the column vector. The "h21qg" command computes the optimal $\mathcal{H}^2$ controller K_{opt} and the associated closed loop system from the exogenous signals w to the regulated signals z. An eigenvalue-eigenvector method is the default method used to solve the two relevant algebraic Riccati equations. A Schur method—based on Schur's unitary transformation of a matrix to upper triangular form—may be used by including the "schur" option. The results are stored in the tree vectors `ss_k` and `ss_twz`, respectively. The 'branch' command is then used to retrieve the state space representation for K_{opt} from the tree vector `ss_k`. ■

Comment 30.12 (Relationship to LQG, Stability Robustness Margins)

Theorem 30.1 shows that the optimal $\mathcal{H}^2$ output feedback controller is identical in structure to that found in classical LQG problems. While certain LQR, KBF, and LQG/LTR problem formulations do result in feedback loops possessing stability robustness margins, LQG controllers need not possess margins [3].

The same is true for $\mathcal{H}^2$ output feedback designs. We will show how the $\mathcal{H}^2$ framework presented can be manipulated to solve LQG/LTR problems which yield model-based controllers with desirable stability robustness margins—comparable to those found in feedback designs resulting from suitably formulated LQR and KBF problems (e.g., infinite upward gain margin, at least 6 dB downward gain margin, at least ±60° phase margin). ■

The following example shows how weighted $\mathcal{H}^2$ mixed sensitivity optimization may be used to design a controller for an unstable system with a time delay.

Example 30.2 (Weighted $\mathcal{H}^2$ Mixed Sensitivity Design for Unstable System with Time Delay)

In this example, we consider an unstable system with a time delay $\Delta = 0.05$ s (50 ms). The system is modeled (approximately) as follows:

$$P \approx \frac{1}{s-1}\left[\frac{2/\Delta - s}{2/\Delta + s}\right] = \frac{1}{s-1}\left[\frac{40-s}{40+s}\right] \tag{30.87}$$

Design Specifications. The objective is to design a controller K that satisfies the following closed loop specifications: (1) closed loop stability, (2) sensitivity below −60 dB for all frequencies below 0.1 rad/s, (3) sensitivity gain crossover between 2 and 3 rad/s, (4) peak sensitivity below 5 dB, (5) peak complementary sensitivity below 10 dB.

Weighted $\mathcal{H}^2$ Mixed Sensitivity Problem. To achieve the above specifications, we formulated a weighted $\mathcal{H}^2$ mixed sensitivity problem—with a weighting W_1 on the sensitivity S and a weighting W_2 on KS; i.e.,

$$\min_K \|T_{wz}\|_{\mathcal{H}^2} = \min_K \left\| \begin{bmatrix} W_1 S \\ W_2 KS \end{bmatrix} \right\|_{\mathcal{H}^2} \tag{30.88}$$

The weighting functions used were as follows:

$$W_1 = \frac{k_1}{s+p_1} = \frac{10}{s+0.01} \tag{30.89}$$

$$W_2 = \frac{k_2(s+z_2)}{s+p_2} = \frac{0.1(s+40)}{s+2} \tag{30.90}$$

W_1 penalizes the sensitivity S heavily at low frequencies (e.g., below 0.001 rad/s). Above 0.1 rad/s, W_1 is small and W_2 penalizes KS (with magnitude greater than unity) until about 4 rad/s. Since the solution of our $\mathcal{H}^2$ optimization depends in a very complex manner on the parameters that define W_1 and W_2, it should be no surprise that it took a while to determine suitable parameters.

Construction of Generalized Plant. The generalized plant G was assembled using SIMULINK and the "linmod" command. The resulting two-port state space representation is as follows:

$$G = \left[\begin{array}{c|c} 0 & -W_1P \\ 0 & W_2 \\ \hline 1 & -P \end{array}\right] = \left[\begin{array}{c|c|c} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ \hline C_2 & D_{21} & D_{22} \end{array}\right] = \left[\begin{array}{cccccc} -0.01 & 0 & -40 & 1 & 1 & 0 \\ 0 & -2 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 40 & -39 & 0 & 1 \\ \hline 10 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0.1 \\ \hline 0 & 0 & -40 & 1 & 1 & 0 \end{array}\right] \tag{30.91}$$

Computation of $\mathcal{H}^2$ Optimal Controller. The "mksys" command was used to pack the above two-port state space into a tree vector data structure. The "h2lqg" command was then used to obtain the optimal controller. Note that the generalized plant is 4th order (two for plant P, one for sensitivity weighting W_1, one for control weighting W_2). The optimal controller:

$$K_{\text{opt}} = \frac{191.0813(s+40)(s+2)(s+0.526)}{(s+1.915)(s+0.01)(s^2+84.15s+2133)} \tag{30.92}$$

is also 4th order—the order of the generalized plant G. The pole at $s = -0.01$ is an approximate integrator—a consequence of the heavy weighting that W_1 places on the sensitivity at low frequencies.

Closed Loop Analysis. The resulting closed loop poles (two plant $P = G_{22}$, four from controller K_{opt}) are as follows:

$$s = -1, -2.0786 \pm j0.8302, -40, -40, -39.9216 \tag{30.93}$$

The resulting sensitivity, KS, and complementary sensitivity frequency responses are shown in Figs.30.4–30.6, respectively. The figures show that all of the design specifications are met (or nearly met). The peak sensitivity is about 4.855 dB. The peak complementary sensitivity is about 8.71 dB. The KS response shows the impact of the compensators' lead between 0.1 and 10 rad/s.

Computation of Minimum $\mathcal{H}^2$ Norm. The minimum two-norm was computed using the following MATLAB command sequence:

```
lc = lyap(acl, bcl*bcl)
minnorm = sqrt( trace( cc1*c*c'))
```

The minimum two-norm was found to be 9.0648. ■

The following simple example illustrates how the $\mathcal{H}^2$ output feedback problem solution can be used to solve classical LQG problems.

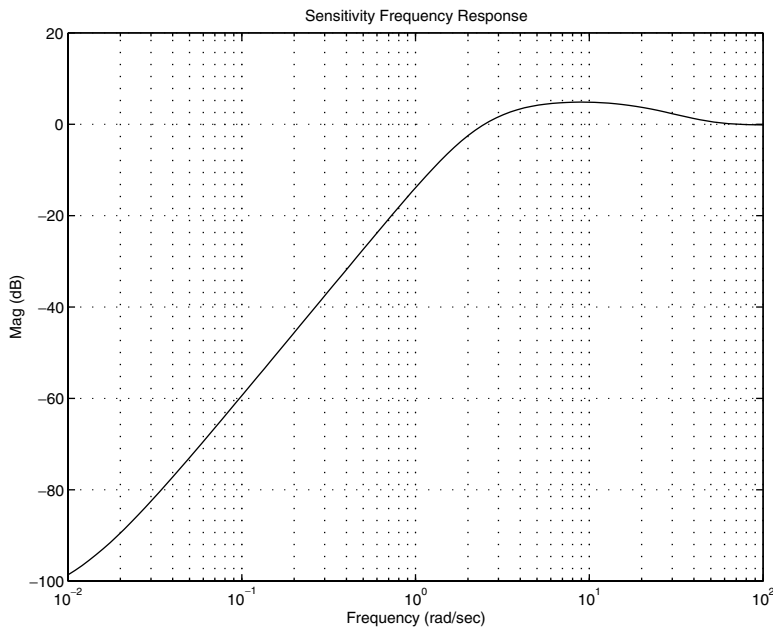



FIGURE 30.4 $\mathcal{H}^2$ Design sensitivity frequency response.

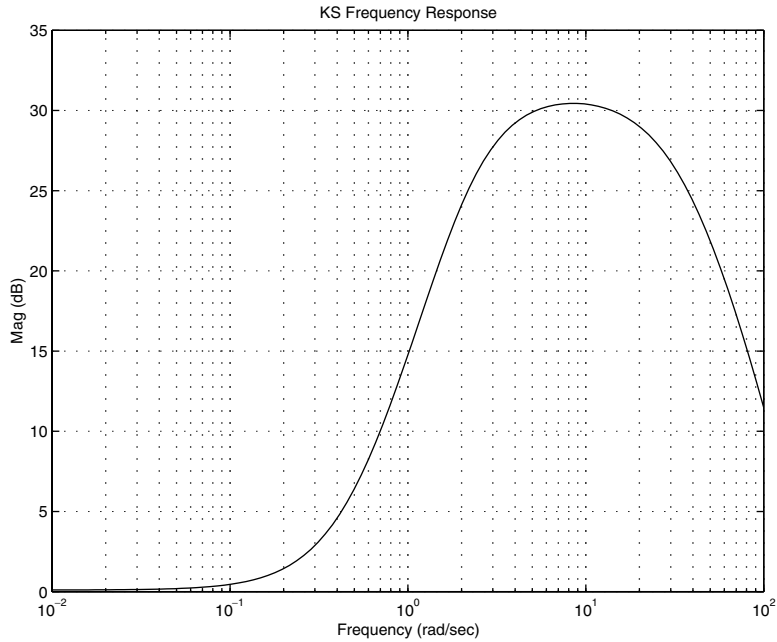



FIGURE 30.5 $\mathcal{H}^2$ Design *KS* frequency response.

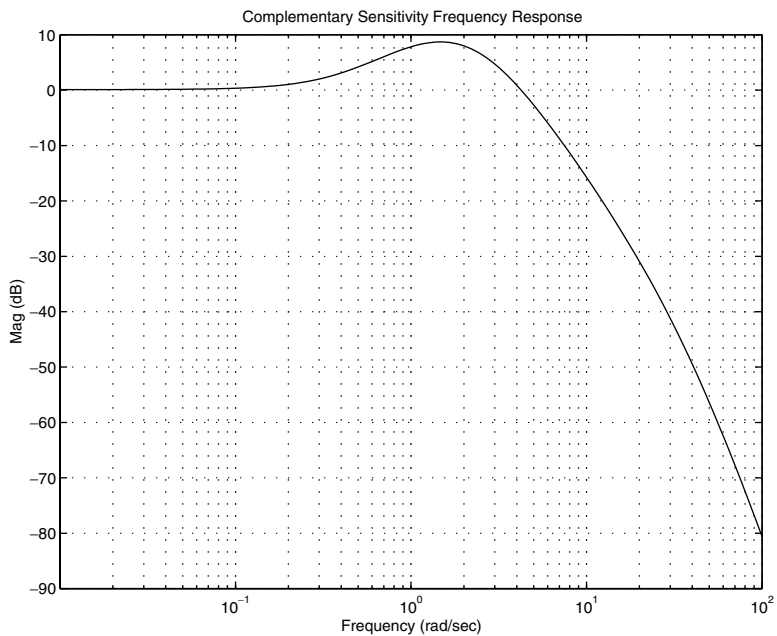



FIGURE 30.6 $\mathcal{H}^2$ Design complementary sensitivity frequency response.

Example 30.3 (LQG/LTR Design for First Order Unstable Missile Model)

We consider an unstable missile described by a simple first order model with state x (pitch attitude), control input u (fin elevator deflection), process noise $w_1 = \xi$ (angular wind gust), and sensor noise $w_2 = \theta$.

It is assumed that the missile's center of gravity (c.g.) is aft of its center of pressure (c.p.)—where lift is concentrated. This assumption results in a missile pitch instability. It is also assumed that the missile's

moment of inertia about its c.g. is very small. This assumption leads to a simple first order model. The missile's angular velocity $\dot{x}$ is assumed to be proportional to its attitude x and the process noise $w_1 = \xi$. Regulated signals $z = [z_1 \ z_2]^T$ include the vehicle's pitch attitude $z_1 = x$ and a weighted control input $z_2 = \sqrt{\rho}u$. Here, $\rho > 0$ is a design parameter to be selected below. The vehicle's pitch attitude is measured. The pitch attitude measurement y includes additive sensor noise $w_2 = \theta$.

Missile Model. The (generalized) missile model is given as follows:

$$\dot{x} = x + \xi + u \tag{30.94}$$

$$z = \begin{bmatrix} x \\ \sqrt{\rho}u \end{bmatrix} \tag{30.95}$$

$$y = x + \sqrt{\mu}\theta \tag{30.96}$$

where $\mu > 0$ is design parameter to be selected below.

Design Specifications. The goal is to design a real-rational proper model-based $\mathcal{H}^2$ optimal compensator (i.e., minimizes $\|T_{wz}\|_{\mathcal{H}^2}$) which results in a stable closed loop system with a dominant closed loop pole at $s = -5$ (settling time $t_s \approx 1$ s).

Construction of Generalized Plant. The above model may be rewritten as follows:

$$\dot{x} = x + \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \xi \\ \theta \end{bmatrix} + u \tag{30.97}$$

$$z = \begin{bmatrix} 1 \\ 0 \end{bmatrix} x + 0_{2\times 2} \begin{bmatrix} \xi \\ \theta \end{bmatrix} + \begin{bmatrix} 0 \\ \sqrt{\sigma} \end{bmatrix} u \tag{30.98}$$

$$y = x + \begin{bmatrix} 0 & \sqrt{\mu} \end{bmatrix} \begin{bmatrix} \xi \\ \theta \end{bmatrix} + 0_{1\times 1}u \tag{30.99}$$

From this, it follows that

$$A = 1, \qquad B_1 = \begin{bmatrix} 1 & 0 \end{bmatrix}, \qquad B_2 = 1 \tag{30.100}$$

$$C_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad D_{11} = 0_{2\times 2}, \qquad D_{12} = \begin{bmatrix} 0 \\ \sqrt{\rho} \end{bmatrix} \tag{30.101}$$

$$C_2 = 1, \qquad D_{21} = \begin{bmatrix} 0 & \sqrt{\mu} \end{bmatrix}, \qquad D_{22} = 0_{1\times 1} \tag{30.102}$$

$\mathcal{H}^2$ Problem Assumptions. We now check each of the $\mathcal{H}^2$ output feedback problem assumptions, as stated in Assumption 30.2. From the above data, it follows that $D_{11} = 0_{2\times 2}$, $D_{22} = 0_{1\times 1}$, and (A, B_2, C_2) is stabilizable and detectable, and

$$R = D_{12}^T D_{12} = \rho > 0 \tag{30.103}$$

$$\Theta = D_{21}^T D_{21} = \mu > 0 \tag{30.104}$$

Since

$$D_{12}^T C_1 = 0 \tag{30.105}$$

$$B_1 D_{21}^T = 0 \tag{30.106}$$

the imaginary axis rank conditions involving (A, B_2, C_1, D_{12}) and (A, B_1, C_2, D_{21}) in Assumption 30.2 become equivalent to (A, C_1) having no imaginary unobservable modes and (A, B_1) having no imaginary uncontrollable modes. These are clearly satisfied since $A = 1$ has no imaginary modes. Given this, it follows that all of the $\mathcal{H}^2$ output feedback problem assumptions in Assumption 30.2 are satisfied.

Plant. Finally, we note that the so-called plant (or missile) transfer function $P = G_{22}$ is given by

$$P = G_{22} = C_2(sI - A)^{-1}B_2 \tag{30.107}$$

$$= \frac{1}{s-1} \tag{30.108}$$

G_{22} is unstable with a right half plane pole at s = 1. G_{22} is also minimum phase (i.e., no zeros in Res > 0).

Filter Gain Matrix H_f. Since $B_1 D_{21}^T = 0$, the associated FARE is given by

$$AY + YA^T + B_1 B_1^T - YC_2^T \Theta^{-1} C_2 Y = Y + Y + 1 - \frac{1}{\mu} Y^2 = 0 \tag{30.109}$$

or

$$Y^2 - 2\mu Y - \mu = 0 \tag{30.110}$$

Application of the quadratic formula and selecting the positive (stabilizing) root yields:

$$Y = \mu + \sqrt{\mu^2 + \mu} \tag{30.111}$$

This yields the following filter gain matrix:

$$H_f = YC_2^T \Theta^{-1} = 1 + \sqrt{1 + \frac{1}{\mu}} \tag{30.112}$$

We now select μ to achieve the given dominant pole specification:

$$A - H_f C_2 = 1 - 1 - \sqrt{1 + \frac{1}{\mu}} = -5 \tag{30.113}$$

This yields

$$\mu = \frac{1}{24} \tag{30.114}$$

The associated KBF open loop transfer function is given by

$$G_{KF} = -C_2(sI - A)^{-1} H_f \tag{30.115}$$

$$= \frac{-6}{s-1} \tag{30.116}$$

We will see below that this will be the approximate open loop transfer function PK_{opt} for the final design. In this sense, G_{KF} represents our target open loop transfer function.

Control Gain Matrix G_c. Since $D_{12}^T C_1 = 0$, the associated CARE is given by

$$A^T X + XA + C_1^T C_1 - XB_2 R^{-1} B_2 X = X + X + 1 - \frac{1}{\rho} X^2 = 0 \tag{30.117}$$

or

$$X^2 - 2\rho X - \rho = 0 \tag{30.118}$$

Application of the quadratic formula and selecting the positive (stabilizing) root yields:

$$X = \rho + \sqrt{\rho^2 + \rho} \tag{30.119}$$

This yields the following control gain matrix:

$$G_c = R^{-1} B_2^T X = 1 + \sqrt{1 + \frac{1}{\rho}} \tag{30.120}$$

This results in a closed loop (regulator) pole at

$$A - B_2 G_c = 1 - 1 - \sqrt{1 + \frac{1}{\rho}} = -\sqrt{1 + \frac{1}{\rho}} \tag{30.121}$$

Note that for large ρ (referred to as expensive control in LQR problems) we have a closed loop pole at $s = -1$, at the left half plane reflection of the plant pole at $s = 1$. We will select the design parameter ρ to be small (referred to as cheap control in LQR problems) so that this closed loop (regulator) pole $s \approx -1/\sqrt{\rho}$ is fast and the closed loop filter pole at $s = -5$ at is the dominant closed loop pole.

$\mathcal{H}^2$ Optimal Output Feedback Model-Based Compensator. The resulting $\mathcal{H}^2$ optimal output feedback model-based compensator is given by

$$K_{\text{opt}} = \left[\begin{array}{c|c} A - B_2 G_c - H_f C_2 & H_f \\ \hline -G_c & 0_{n_u \times n_y} \end{array}\right] \tag{30.122}$$

where

$$A - B_2 G_c - H_f C_2 = 1 - 1 - \sqrt{1 + \frac{1}{\rho}} - 1 - \sqrt{1 + \frac{1}{\mu}} \tag{30.123}$$

$$= 1 - 1 - \sqrt{1 + \frac{1}{\rho}} - 1 - \sqrt{1 + 24} \tag{30.124}$$

$$= -6 - \sqrt{1 + \frac{1}{\rho}} \tag{30.125}$$

$$H_f = 1 + \sqrt{1 + \frac{1}{\mu}} \tag{30.126}$$

$$= 1 + \sqrt{1 + 24} \tag{30.127}$$

$$= 6 \tag{30.128}$$

$$G_c = 1 + \sqrt{1 + \frac{1}{\rho}} \tag{30.129}$$

Given this, the compensator transfer function is given by

$$K_{\text{opt}} = -G_c(sI - A + B_2G_c + H_fC_2)^{-1}H_f \tag{30.130}$$

$$= \frac{-6(1+\sqrt{1+1/\rho})}{s+6+\sqrt{1+1/\rho}} \tag{30.131}$$

For small ρ (cheap control), this yields

$$K_{\text{opt}} \approx \frac{-6(1/\sqrt{\rho})}{s+1/\sqrt{\rho}} \tag{30.132}$$

Open Loop Transfer Function. The associated open loop transfer function is given by

$$PK_{\text{opt}} = -C_2(sI-A)^{-1}B_2 \quad G_c(sI - A + B_2G_c + H_fC_2)^{-1}H_f \tag{30.133}$$

$$= \frac{1}{s-1}\left[\frac{-6(1+\sqrt{1+1/\rho})}{s+6+\sqrt{1+1/\rho}}\right] \tag{30.134}$$

For small ρ (cheap control), this becomes

$$PK_{\text{opt}} \approx \frac{1}{s-1}\left[\frac{-6(1/\sqrt{\rho})}{s+1/\sqrt{\rho}}\right] \tag{30.135}$$

Loop Transfer Recovery (*LTR*). From this, we see that as control weighting parameter ρ approaches zero (cheap control), the open loop transfer function approaches the KBF open loop transfer function G_{KF}; i.e.,

$$\lim_{\rho \to 0^+} G_{22}K_{\text{opt}} = \frac{-6}{s-1} \tag{30.136}$$

$$= G_{\text{KF}} \tag{30.137}$$

This shows that as ρ approaches zero (cheap control), the actual open loop transfer function PK_{opt} approaches the target open loop transfer function G_{KF}. The above procedure of recovering a target open loop transfer function (with desirable closed loop properties) using an LQG controller is called *LQG with loop transfer recovery or LQG/LTR.*

Selection of Far Away Closed Loop Regulator Pole. For small ρ, the closed loop system is stable with closed loop poles at $s = -5$ and $s \approx -1/\sqrt{\rho}$. A good selection for ρ might be $\rho = 1/2500$. This results in a fast closed loop pole at $s \approx -50$ and makes the closed loop filter pole at $s = -5$ the dominant closed loop pole, as required.

Stability Robustness Margins. It is well known that $\mathcal{H}^2$ and LQG designs need not possess good stability robustness margins. In fact, they can be arbitrarily bad [3]. LQG/LTR designs for minimum phase plants (such as ours: P = 1/(s − 1) have guaranteed stability robustness margins. LQG/LTR designs provide margins that approach those associated with LQR and KBF designs; i.e., infinite upward gain margin, at least 6 dB downward gain margin, and at least ±60° phase margin. Our final LQG/LTR design

$$PK_{\text{opt}} = \frac{-6}{s-1}\left[\frac{50}{s+50}\right] \tag{30.138}$$

offers an infinite upward gain margin and a downward gain margin of 1/6 (−15.56 dB). The resulting unity gain crossover frequency is $\omega_g = \sqrt{35} = 5.92$ rad/s and the associated phase margin is about 99.59°. Not bad.

The following example extends the LQG/LTR ideas presented in Example 30.3 to the general MIMO setting—enabling the design of feedback loops (with nominal robustness margins) via $\mathcal{H}^2$ optimization. ■

Example 30.4 (MIMO LQG and LQG/LTR Control Design Via $\mathcal{H}^2$ Optimization)

We consider a MIMO plant P defined by the state space representation

$$\dot{x} = Ax + Bu \tag{30.139}$$

$$y = Cx \tag{30.140}$$

It is assumed that the plant $P = [A, B, C]$ is stabilizable and detectable.

The goal is to demonstrate how the $\mathcal{H}^2$ optimal output feedback solution that has been presented may be used to solve MIMO LQG control problems. We specifically would like to present a method which lends itself to the concept of LTR—whereby we use a model-based LQG controller to recover a target loop transfer function matrix with desirable closed loop properties. Our motivation is not optimal stochastic LQG control problems; it is the design of control laws with desirable closed loop properties.

Construction of Generalized Plant G. With our final objective being a model-based compensator defined by a control gain matrix G_c and a filter gain matrix H_f, we consider the following generalized plant:

$$\dot{x} = Ax + L\xi + Bu \tag{30.141}$$

$$z = \left[\frac{Mx}{\sqrt{\rho} I_{n_u \times n_u}}\right] \tag{30.142}$$

$$y = Cx + \sqrt{\mu}\theta \tag{30.143}$$

where u is the control, x is the (generalized) plant state, $w_1 = \xi$ represents process noise in the state equation, $w_2 = \theta$ represents sensor noise in the measurement equation, $A \in \mathcal{R}^{n \times n}$, $L \in \mathcal{R}^{n \times n_u}$, $B \in \mathcal{R}^{n \times n_u}$, $M \in \mathcal{R}^{n_y \times n}$, $C \in \mathcal{R}^{n_y \times n}$, $n_y = n_u$, $\rho > 0$, $\mu > 0$.

Design Parameter Assumptions. It is assumed that either:

(A, L) has no imaginary uncontrollable modes and (A, M) is detectable, or
(A, L) is stabilizable and (A, M) has no imaginary unobservable modes.

Here, L, M, μ, and ρ should be viewed as "design parameters" that are selected in order to obtain control and filter gain matrices G_c and H_f such that the resulting model-based compensator exhibits desirable closed loop properties.

Two-Port State Space Representation for Generalized Plant G. The above model may be rewritten in two-port state space form as follows

$$\left[\begin{array}{c}\dot{x}\\ \hline z\\ \hline y\end{array}\right] = \left[\begin{array}{c|c|c} A & \begin{bmatrix} L & 0_{n \times n_y}\end{bmatrix} & B \\ \hline \begin{bmatrix} M \\ 0_{n_u \times n}\end{bmatrix} & \begin{bmatrix} 0_{n_y \times n_u} & 0_{n_y \times n_y} \\ 0_{n_u \times n_u} & 0_{n_u \times n_y}\end{bmatrix} & \begin{bmatrix} 0_{n_y \times n_u} \\ \sqrt{\rho} I_{n_u \times n_u}\end{bmatrix} \\ \hline C & \begin{bmatrix} 0_{n_y \times n_u} & \sqrt{\mu} I_{n_y \times n_y}\end{bmatrix} & \end{array}\right] \left[\begin{array}{c} x \\ \hline \xi \\ \theta \\ \hline u \end{array}\right] \tag{30.144}$$

Check on $\mathcal{H}^2$ Output Feedback Assumptions. We now make sure that all of the $\mathcal{H}^2$ output feedback problem assumptions in Assumption 30.2 are satisfied.

- *Plant $P = G_{22}$ Assumptions.* Since the plant $P = G_{22} = [\ A, B, C\ ]$ is stabilizable and detectable, it follows that $(A, B_2 = B, C_2 = C)$ is stabilizable and detectable.
- *Regulator Assumptions.* Since

$$D_{12} = \begin{bmatrix} 0_{n_y \times n_u} \\ \sqrt{\rho} I_{n_u \times n_u} \end{bmatrix}$$

has full column rank, it follows that the control weighting matrix $R = D_{12}^T D_{12} = \rho I_{n_u \times n_u} > 0$ is nonsingular.

Since $D_{12}^T C_1 = 0$, it follows that the imaginary axis (column) rank condition involving (A, B_2, C_1, D_{12}) in Assumption 30.2 is equivalent to $(A - B_2 R^{-1} D_{12}^T C_1, (I - D_{12} R^{-1} D_{12}^T) C_1) = (A, C_1)$ having no unobservable imaginary modes. Since (A, M) is either detectable or has no imaginary unobservable modes, it follows that

$$A, C_1 = \begin{bmatrix} M \\ 0_{n_u \times n} \end{bmatrix}$$

has no unobservable imaginary modes. The associated Hamiltonian H_{con} will, therefore, yield a Riccati solution and control gain matrix G_c such that $A - BG_c$ is stable.
- *Filter Assumptions.* Since $D_{21} = [0_{n_y \times n_u} \quad \sqrt{\mu} I_{n_y \times n_y}]$ has full row rank, it follows that the measurement weighting matrix $\Theta = D_{21} D_{21}^T = \mu I_{n_y \times n_y} > 0$ is nonsingular.

Since $B_1 D_{21}^T = 0$, it follows that the imaginary axis (row) rank condition involving (A, B_1, C_2, D_{21}) in Assumption 30.2 is equivalent to $(A - B_1 D_{21}^T \Theta^{-1} C_2, B_1(I - D_{21} \Theta^{-1} D_{21})) = (A, B_1)$ having no uncontrollable imaginary modes. Since (A, L) is either stabilizable or has no uncontrollable imaginary modes, it follows that $(A, B_1 = [L \quad 0_{n \times n_y}])$ has no uncontrollable imaginary modes. The associated Hamiltonian H_{fil} will therefore yield a Riccati solution and filter gain matrix H_f such that $A - H_f C$ is stable.

Given the above, it follows that all of the $\mathcal{H}^2$ output feedback problem assumptions in Assumption 30.2 are satisfied.

Control Gain Matrix. It follows that the control gain matrix G_c is given by

$$G_c = R^{-1} B^T X \tag{30.145}$$

where $X \geq 0$ is the unique (at least) positive semi-definite solution of the CARE:

$$A^T X + XA + C^T C - XBR^{-1} BX = 0 \tag{30.146}$$

Moreover, $A - BG$ is stable.

Filter Gain Matrix. It follows that the filter gain matrix H_f is given by

$$H_f = YC^T \Theta^{-1} \tag{30.147}$$

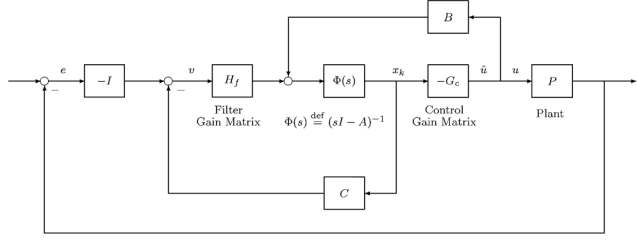



FIGURE 30.7 Negative feedback loop with LQG model-based compensator and plant.

where $Y > 0$ is the unique (at least) positive semi-definite solution of the FARE:

$$AY + YA^T + LL^T - YC^T\Theta^{-1}CY = 0 \tag{30.148}$$

Moreover, $A - HC$ is stable.

$\mathcal{H}^2$ Optimal (LQG) Compensator. The $\mathcal{H}^2$ optimal compensator that minimizes the $\mathcal{H}^2$ norm of the transfer function matrix from the exogenous signals $w = \begin{bmatrix} \xi \\ \theta \end{bmatrix}$ to the regulated signals $z = \begin{bmatrix} C_x \\ \sqrt{\rho} I_{n_u \times n_u} \end{bmatrix}$ is then given by

$$K_{\text{opt}} = \left[\begin{array}{c|c} A - BG_c - H_f C & H_f \\ \hline G_c & 0_{n_u \times n} \end{array}\right] \tag{30.149}$$

Note that the minus sign on G_c (lower left hand entry of K_{opt}) has been removed in anticipation of the negative feedback system implementation shown in Fig. 30.7. By the separation principle, the closed loop poles are the eigenvalues of $A - BG_c$ and $A - H_f C$.

Stability Robustness Margins. It should be emphasized that the resulting controller K_{opt}, although stabilizing, may possess arbitrarily bad stability robustness margins [3]. This is despite the fact that the associated regulator loop

$$G_{LQ} = G_c(sI - A)^{-1}B \tag{30.150}$$

and filter loop

$$G_{\text{KF}} = C(sI - A)^{-1}H_f \tag{30.151}$$

when viewed as MIMO open loop transfer function matrices within their own negative feedback loops, possess the following well-known stability robustness margins: infinite upward gain margin, at least 6 dB downward gain margin, and at least ±60° phase margin. This gives rise to the following natural question:

Is there a way that we can select the control gain matrix G_c and the filter gain matrix H_f so that the resulting model-based compensator K_{opt} results in a feedback loop which possesses the above nice margins?

Fortunately, the answer to this is a definitive yes! Two methods which result in comparable stability margins at the plant input or at the plant output (but not both simultaneously) are now presented.

Loop Transfer Recovery (LTR) Methods. The approach we take to achieve a feedback design with good stability margins is as follows. The process involves two steps.

1. *Target Loop Design.* The first step is to design a target open loop transfer function matrix that possesses desired closed loop properties. The target loop may be associated with the plant output. If so, we denote it L_o. In such a case, L_o represents our desired PK_{opt}. If associated with the plant input, we denote it L_i. In such a case, L_i represents our desired K_{opt}. (In general, $PK_{\text{opt}}P \neq K_{\text{opt}}P$.)

2. *Target Loop Recovery Via Model-Based Compensator.* The second step is to use a model-based compensator $K_{opt} = [A - BG_c - H_f C, H, G]$ to recover the target loop (either L_o or L_i).

   If we want to recover L_o (i.e., good properties at that plant output), then we want $PK_{opt} \approx L_o$. This is called loop transfer recovery at the plant output (LTRO).

   If we want to recover L_i (i.e., good properties at that plant input), then we want $K_{opt}P \approx L_i$. This is called loop transfer recovery at the plant input (LTRI).

   Note: In general, the properties associated with breaking the loop at the plant output (properties of PK_{opt}) are different (perhaps very different) from those associated with breaking the loop at the plant input (properties of $K_{opt}P$). It is usually very difficult for PK_{opt} and $K_{opt}P$ to both possess great properties (e.g., margins, etc.). Typically, a designer must trade off nice properties at the plant output for nice properties at the plant input, or vice versa.

$\mathcal{H}^2$-based methods for LTRO and LTRI are now presented.

- Loop Transfer Recovery at Plant Output (LTRO).
  1. *Design of Target Loop L_o.* The first step is to design a target loop $L_o = C(sI - A)^{-1}H_f$ with desirable closed loop properties (e.g., stability, sensitivity, complementary sensitivity, stability robustness margins, etc). This may be done using any method! (Any method you feel comfortable enough with.)

     One procedure that results in good properties at the plant output is based on *KBF* methods. The idea is to select the design (shaping) matrix L so that the singular values $G_{FOL} = C(sI - A)^{-1}L$ look nice; e.g., large minimum singular value at low frequencies, small maximum singular value at high frequencies, singular values cross 0 dB with slopes of –20 dB/dec, etc.

     We then solve the FARE with A, L, C, $\Theta = \mu I_{n_y \times n_y}$—using $\mu > 0$ to adjust the bandwidth of our target loop $L_o = G_{KF} = C(sI - A)^{-1}H_f$. A smaller (larger) μ results in a larger (smaller) bandwidth.

     Guidelines for Shaping of Target Loop $L_o = G_{KF}$.

- The so-called Kalman Frequency Domain Equality (KFDE) guides our loop shaping:

$$[I + G_{KF}(j\omega)][I + G_{KF}(j\omega)]^H = I + \left[\frac{1}{\sqrt{\mu}}G_{FOL}(j\omega)\right]\left[\frac{1}{\sqrt{\mu}}G_{FOL}(j\omega)\right]^H \tag{30.152}$$

From this, it follows that

$$\sigma_i[I + G_{KF}(j\omega)] = \sqrt{1 + \frac{1}{\mu}\sigma_i^2[G_{FOL}(j\omega)]} \tag{30.153}$$

This suggests that by shaping G_{FOL}, we can shape the target loop $L_o = G_{KF}$. Specifically, if G_{FOL} is large at low frequencies, then we expect

$$G_{KF}(j\omega) \approx \frac{1}{\sqrt{\mu}}G_{FOL}(j\omega) \tag{30.154}$$

at low frequencies. This shows that the matrix L should be used for shaping the target loop $L_o = G_{KF}$ while $\mu > 0$ is used to adjust the target loop bandwidth—decreasing/increasing μ to raise/lower the target loop bandwidth. The resulting loop $L_o = G_{KF}$ is guaranteed to possess nice closed loop properties as described below.

- The above singular value relation implies that

$$\sigma_{min}[I + G_{KF}(j\omega)] \geq 1 \tag{30.155}$$

for all ω. This, in turn, implies that the associated sensitivity singular values satisfy

$$\sigma_{\max}[S_{KF}(j\omega)] = \frac{1}{\sigma_{\min}[S_{KF}(j\omega)]^{-1}} \leq 1 \quad (0\text{dB}) \tag{30.156}$$

for all ω, where

$$S_{KF}(j\omega) = [I + G_{KF}(j\omega)]^{-1} \tag{30.157}$$

- From the above sensitivity singular value relationship, we obtain the follow celebrated KBF loop margins:

  infinite upward gain margin,
  at least $\frac{1}{2}$ (6 dB) downward gain margin,
  at least ±60° phase margin.

  The above gain margins apply to simultaneous and independent gain perturbations when the loop is broken at the output. The same holds for the above phase margins. The above margins are NOT guaranteed for simultaneous gain and phase perturbations. It should be noted that these margins can be easily motivated using elementary SISO Nyquist stability arguments [2,8].
- From the above sensitivity singular value relations, we obtain the following complementary sensitivity singular value relationship:

$$\sigma_{\max}[T_{KF}(j\omega)] = \sigma_{\max}[I - S_{KF}(j\omega)] \leq 1 + \sigma_{\max}[S_{KF}(j\omega)] \leq 2 \quad (6\text{dB}) \tag{30.158}$$

for all ω, where

$$T_{KF} = I - S_{KF} = G_{KF}[1 + G_{KF}]^{-1} \tag{30.159}$$

2. *Recovery of Target Loop L_o Using Model-Based Compensator.* The second step is to use a model-based compensator $K_{opt} = [\ A - BG_c - H_fC, H_f, G_c\ ]$ where G_c is found by solving the CARE with A, B, $M = C$, $R = \rho I_{n_u \times n_u}$ with ρ a small positive scalar. Since ρ is small, we call this a cheap control problem.
   - If the plant $P = [A, B, C]$ is minimum phase, then it can be shown that

$$\lim_{\rho \to 0^+} X = 0 \tag{30.160}$$

$$\lim_{\rho \to 0^+} \sqrt{\rho} G_c = WC \tag{30.161}$$

   for some orthonormal W (i.e., $W^TW = WW^T = I$)

$$\lim_{\rho \to 0^+} PK_{opt} = L_o \tag{30.162}$$

   In such a case, $PK_{opt} \approx L_o$ for small ρ and hence PK_{opt} will possess stability margins that are close to those of L_o (at the plant output)—whatever method was used to design L_o. It must be noted that the minimum phase condition on the plant P is a suficient condition. It is not necessary. Moreover, G_c need not be computed using a CARE. In fact, any G_c which (1) satisfies a limiting condition $\lim_{\rho \to 0^+} \sqrt{\rho} G_c = WC$ for some invertible matrix W and which (2) ensures that $A - BG_c$ is stable (for small ρ), will result in LTR at the plant output. This result

is a consequence of the structure of model-based compensators and has nothing to do with optimal control and filtering problems.

- Assuming that a limiting condition $\lim_{\rho \to 0^+} \sqrt{\rho} G_c = WC$ holds for some invertible matrix W, loop transfer recovery of the target loop $L_o = C(sI - A)^{-1} H_f$ may be proven as follows: For small ρ we have

$$G_c \approx \frac{WC}{\sqrt{\rho}} \tag{30.163}$$

which gives yields the following:

$$PK_{\text{opt}} = PG_c(sI - A + BG_c + H_f C)^{-1} H_f \tag{30.164}$$

$$\approx P\frac{WC}{\sqrt{\rho}}\left(sI - A + B\frac{WC}{\sqrt{\rho}}\right)^{-1} H_f \tag{30.165}$$

$$\approx P\frac{WC}{\sqrt{\rho}}(sI - A)^{-1}\left[I + B\frac{WC}{\sqrt{\rho}}(sI - A)^{-1}\right]^{-1} H_f \tag{30.166}$$

$$\approx P\frac{WC(sI - A)^{-1}}{\sqrt{\rho}}\left[I + B\frac{WC(sI - A)^{-1}}{\sqrt{\rho}}\right]^{-1} H_f \tag{30.167}$$

$$\approx P\left[I + \frac{WC(sI - A)^{-1}}{\sqrt{\rho}}B\right]^{-1}\frac{WC(sI - A)^{-1}}{\sqrt{\rho}}H_f \tag{30.168}$$

$$\approx P\left[\frac{WP}{\sqrt{\rho}}\right]^{-1} W\frac{C(sI - A)^{-1}H_f}{\sqrt{\rho}} \tag{30.169}$$

$$\approx C(sI - A)^{-1} H_f = L_o \tag{30.170}$$

The central idea (underneath the algebra) is that as ρ goes to zero, the C feedback path within the compensator $K_{\text{opt}} = [A - BG_c - H_f C, G_c, H_f]$ is broken (see Fig.30.7) and the nice properties that hold at the so-called innovations ν (e.g., open loop transfer function matrix at ν is $L_o = C(sI - A)^{-1}H_f$) in Fig. 30.7 get transferred to the error signal e (compensator input, or plant output) within the feedback loop.

- Loop Transfer Recovery at Plant Input (LTRI).

1. *Design of Target Loop L_i.* The first step is to design a target loop $L_i = G_C(sI - A)^{-1}B$ with desirable closed loop properties (e.g., stability, sensitivity, complementary sensitivity, stability robustness margins, etc). This may be done using any method! (Any method you feel comfortable enough with.)

   One procedure that results in good properties at the plant input is based on LQR methods. The idea is to select the design (shaping) matrix M so that the singular values of $G_{\text{OL}} = M(sI - A)^{-1}B$ look nice; e.g., large minimum singular value at low frequencies, small maximum singular value at high frequencies, singular values cross 0 dB with slopes of –20 dB/dec, etc.

   We then solve the CARE with A, B, M, $R = \rho I_{n_u \times n_u}$, using $\rho > 0$ to adjust the bandwidth of our target loop $L_i = G_{\text{LQR}} G_c(sI - A)^{-1}B$. A smaller (larger) ρ results in a larger (smaller) bandwidth.

   Guidelines for Shaping of Target Loop $L_i = G_{\text{LQ}}$.

- The so-called LQ frequency domain equality (LQFDE) guides our loop shaping:

$$[I + G_{\text{LQ}}(j\omega)]^H[I + G_{\text{LQ}}(j\omega)] = I + \left[\frac{1}{\sqrt{\rho}}G_{\text{OL}}(j\omega)\right]^H\left[\frac{1}{\sqrt{\rho}}G_{\text{OL}}(j\omega)\right] \tag{30.171}$$

From this, it follows that

$$\sigma_i[I + G_{LQ}(j\omega)] = \sqrt{I + \frac{1}{\rho}\sigma_i^2[G_{\mathrm{OL}}(j\omega)]} \tag{30.172}$$

This suggests that by shaping G_{OL}, we can shape the target loop $L_i = G_{\mathrm{LQ}}$. Specifically, if G_{OL} is large at low frequencies, then we expect

$$G_{\mathrm{LQ}}(j\omega) \approx \frac{1}{\sqrt{\rho}} G_{\mathrm{OL}}(j\omega) \tag{30.173}$$

at low frequencies. This shows that the matrix M should be used for shaping the target loop $L_i = G_{\mathrm{LQ}}$ while $\rho > 0$ is used to adjust the target loop bandwidth—decreasing/increasing ρ to raise/lower the target loop bandwidth. The resulting loop $L_i = G_{\mathrm{LQ}}$ is guaranteed to possess nice closed loop properties as described below.

- The above singular value relation implies that

$$\sigma_{\min}[I + G_{\mathrm{LQ}}(j\omega)] \geq 1 \tag{30.174}$$

for all ω. This, in turn, implies that the associated sensitivity singular values satisfy

$$\sigma_{\max}[S_{\mathrm{LQ}}(j\omega)] = \frac{1}{\sigma_{\min}[S_{\mathrm{LQ}}(j\omega)^{-1}]} \leq 1 \quad (0\mathrm{dB}) \tag{30.175}$$

for all ω, where

$$S_{\mathrm{LQ}} = [I + G_{\mathrm{LQ}}]^{-1} \tag{30.176}$$

- From the above sensitivity singular value relationship, we obtain the follow celebrated LQR loop margins:

  infinite upward gain margin,
  at least $\frac{1}{2}$(6 dB) downward gain margin,
  at least ±60° phase margin.

  The above gain margins apply to simultaneous and independent gain perturbations when the loop is broken at the input. The same holds for the above phase margins. The above margins are NOT guaranteed for simultaneous gain and phase perturbations. It should be noted that these margins can be easily motivated using elementary SISO Nyquist stability arguments [2,8].
- From the above sensitivity singular value relations, we obtain the following complementary sensitivity singular value relationship:

$$\sigma_{\max}[T_{\mathrm{LQ}}(j\omega)] = \sigma_{\max}[I - S_{\mathrm{LQ}}(j\omega)] \leq 1 + \sigma_{\max}[S_{\mathrm{LQ}}(j\omega)] \leq 2 \quad (6\mathrm{dB}) \tag{30.177}$$

for all ω, where

$$T_{\mathrm{LQ}} = I - S_{\mathrm{LQ}} = G_{\mathrm{LQ}}[1 + G_{\mathrm{LQ}}]^{-1} \tag{30.178}$$

2. Recovery of Target Loop L_i Using Model-Based Compensator. The second step is to use $K_{\mathrm{opt}} = [A - BG_c, -H_fC, H_f, G_c]$ where H_f is found by solving the FARE with A, $L = B$, C, $\Theta = \mu I_{n_y \times n_y}$ with μ a small positive scalar. Since μ is small, we call this an expensive sensor problem.

- If the plant $P = [A, B, C]$ is minimum phase, then it can be shown that

$$\lim_{\mu \to 0^+} Y = 0 \tag{30.179}$$

$$\lim_{\mu \to 0^+} \sqrt{\mu} H_f = BV \tag{30.180}$$

for some orthonormal V (i.e., $V^T V = VV^T = I$)

$$\lim_{\mu \to 0+} K_{\text{opt}} P = L_i \tag{30.181}$$

In such a case, $K_{\text{opt}}\, P \approx L_i$ for small μ and hence $K_{\text{opt}}\, P$ will possess stability margins that are close to those of L_i (at the plant input)—whatever method was used to design L_i.

It must be noted that the minimum phase condition on the plant P is a suficient condition. It is not necessary. Moreover, H_f need not be computed using a FARE. In fact, any H_f which (1) satisfies a limiting condition $\lim_{\mu \to 0^+} \sqrt{\mu} H_f = BV$ for some invertible matrix V and which (2) ensures that $A - H_f C$ is stable (for small μ), will result in LTR at the plant input; i.e., $\lim_{\mu \to 0^+} K_{\text{opt}} P = L_i$. This result is a consequence of the structure of model-based compensators and has nothing to do with optimal control and filtering problems.

- Assuming that a limiting condition $\lim_{\mu \to 0^+} \sqrt{\mu} H_f = BV$ holds for some invertible matrix V, loop transfer recovery of the target loop $L_i = G_c(sI - A)^{-1}$ B may be proven as follows. For small μ we have

$$H_f \approx \frac{BV}{\sqrt{\mu}} \tag{30.182}$$

which gives the following:

$$K_{\text{opt}} P = G_c(sI - A + BG_c + H_f C)^{-1} H_f P \tag{30.183}$$

$$\approx G_c\left(sI - A + BG_c + \frac{BV}{\sqrt{\mu}} C\right)^{-1} \frac{BV}{\sqrt{\mu}} P \tag{30.184}$$

$$\approx G_c\left(sI - A + \frac{BV}{\sqrt{\mu}} C\right)^{-1} \frac{BV}{\sqrt{\mu}} P \tag{30.185}$$

$$\approx G_c(sI - A)^{-1} + \left[I + \frac{BV}{\sqrt{\mu}} C(sI - A)^{-1}\right]^{-1} \frac{BV}{\sqrt{\mu}} P \tag{30.186}$$

$$\approx G_c(sI - A)^{-1} \frac{BV}{\sqrt{\mu}} \left[I + C(sI - A)^{-1} \frac{BV}{\sqrt{\mu}}\right]^{-1} P \tag{30.187}$$

$$\approx G_c(sI - A)^{-1} \frac{BV}{\sqrt{\mu}} \left[C(sI - A)^{-1} \frac{BV}{\sqrt{\mu}}\right]^{-1} P \tag{30.188}$$

$$\approx G_c(sI - A)^{-1} B \frac{V}{\sqrt{\mu}} \left[P \frac{V}{\sqrt{\mu}}\right]^{-1} P \tag{30.189}$$

$$\approx G_c(sI - A)^{-1} B = L_i \tag{30.190}$$

The central idea (underneath the algebra) is that as μ goes to zero, the B feedback path within the compensator $K_{\text{opt}} = [A - BG_c - H_f C, G_c, H_f]$ is broken (see Fig. 30.7) and the nice properties

that hold at $\hat{u}$ (e.g., open loop transfer function matrix at $\hat{u}$ is $L_i = G_c(sI - A)^{-1}B$) in Fig. 30.7 get transferred to the plant input u (compensator output) within the feedback loop. ■

Comment 30.13 (Stability Margins and Peak Sensitivity)

The peak on the sensitivity plot is very important in the design of a feedback system. A large peak, for example, may be due to a closed loop pole near the imaginary axis. This certainly is undesirable. We thus want the peak to be "small." It can be shown that the peak necessarily establishes gain and phase margin bounds.

Suppose that the peak sensitivity is bounded above by $\alpha \geq 1$; i.e., $\sigma_{\max} S(j\omega) < \alpha$ for all ω. It can be shown that the feedback loop then enjoys the following nominal multivariable stability robustness (gain and phase) margin bounds:

$$\uparrow GM > \frac{\alpha}{\alpha - 1} \tag{30.191}$$

$$\downarrow GM < \frac{\alpha}{\alpha + 1} \tag{30.192}$$

$$|PM| > 2\sin^{-1}\left(\frac{1}{2\alpha}\right) \tag{30.193}$$

These bounds may be easily motivated using SISO Nyquist [2,8] ideas as follows.

If

$$|S(j\omega)| < \alpha \tag{30.194}$$

for all ω, then it follows that

$$\frac{1}{\alpha} < |1 + L(j\omega)| \tag{30.195}$$

for all ω. This, however, implies that the Nyquist plot associated with L cannot penetrate a circle centered at -1 with radius $1/\alpha$, left most end point at $-[(\alpha+1)/\alpha]$, and right most end point at $-[(\alpha-1)/\alpha]$. The upward gain margin bound follows from the right most point of the circle. The downward gain margin bound follows from the left most point of the circle. The phase margin bound can be obtained with a little geometry. ■

The following example considers the application of $\mathcal{H}^2$ theory to a robotic manipulator.

Example 30.5 ($\mathcal{H}^2$-LQG/LTR Design for PUMA 560 Robotic Manipulator)

In this example, we show how $\mathcal{H}^2$ optimization may be used to design an LQG/LTR controller for a PUMA 560 robotic manipulator. The manipulator is shown in Fig. 30.8.

A two degree-of-freedom (dof) linear model $P = [A_p, B_p, C_p]$ was used to initiate the design process. Linearizing the PUMA's nonlinear model [9] about the equilibrium point $\theta_1 = 90° \; \theta_2 = 0°$ (both links vertical), results in the following linear model:

$$\dot{x}_p = A_p x_p + B_p u_p \tag{30.196}$$

$$y_p = C_p x_p \tag{30.197}$$

$$u_p = [\tau_1 \quad \tau_2]^T \tag{30.198}$$

$$x_p = [\theta_1 \quad \theta_2 \quad \dot{\theta}_1 \quad \dot{\theta}_2]^T \tag{30.199}$$

$$y_p = [\theta_1 \quad \theta_2]^T \tag{30.200}$$

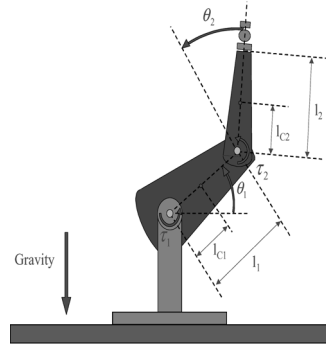



FIGURE 30.8 Two degree-of-freedom PUMA 560 robotic manipulator.

where

$$A_p = \begin{bmatrix} 0.0000 & 0.0000 & 1.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 1.0000 \\ 31.7613 & -33.0086 & 0.0000 & 0.0000 \\ -56.9381 & 187.7089 & 0.0000 & 0.0000 \end{bmatrix} \quad (30.201)$$

$$B_p = \begin{bmatrix} 0.0000 & 0.0000 \\ 0.0000 & 0.0000 \\ 1037.7259 & -3919.6674 \\ -3919.6674 & 2030.8306 \end{bmatrix} \quad (30.202)$$

$$C_p = [I_{2\times 2} \quad 0_{2\times 2}] \quad (30.203)$$

The system poles are $s = \pm 14.1050$, $s = \pm 4.5299$. Eigenvector analysis shows that the fast instability at $s = 14.1050$ is primarily associated with the upper (shorter) link, while the slower instability at $s = 4.5299$ is primarily associated with the lower (longer) link. The system does not possess any natural integrators (i.e., no zero eigenvalues) and, as expected, the singular values $\sigma_i[P(j\omega)]$ are flat at low frequencies (see Fig. 30.9).

Closed Loop Objectives

A controller to be implemented within a negative feedback loop is sought. The closed loop system should exhibit the following properties: (1) closed loop stability, (2) zero steady state error to step reference commands, (3) good low frequency reference command following (step commands followed with little overshoot within 3 s), (4) good low frequency disturbance attenuation, (5) good high frequency noise attenuation, (6) good stability robustness margins at the plant output.

Each step of the control system design process is now described. A central idea is the formation of a so-called design plant P_d from the original plant P. The design plant P_d is what is submitted to our $\mathcal{H}^2$-LQG/LTR design machinery.

***Step 1: Augment Plant P with Integrators to Get Design Plant Pd* = [*A, B, C*]**

In order to guarantee zero steady-state error to step reference commands, we begin by augmenting the plant $P = [A_p, B_p, C_p]$ with integrators—one in each control channel—to form the design plant $P_d = [A, B, C]$; i.e., $P_d = P(I_{2\times 2}/s)$. This is done as follows:

$$A = \begin{bmatrix} 0_{2\times 2} & 0_{2\times 4} \\ B_p & A_p \end{bmatrix} \quad (30.204)$$

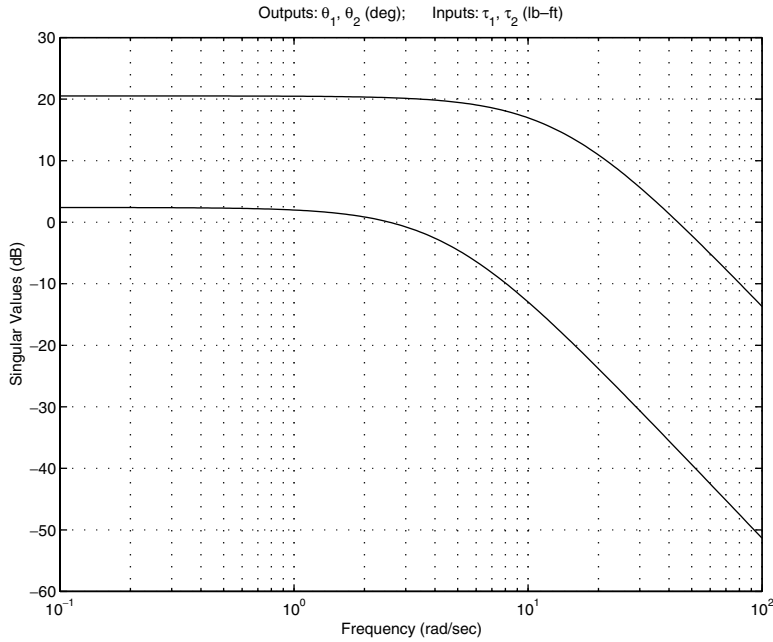



FIGURE 30.9 PUMA 560 robotic manipulator singular values.

$$B = \begin{bmatrix} 1_{2\times 2} \\ 0_{4\times 2} \end{bmatrix} \tag{30.205}$$

$$C = [0_{2\times 2} \quad C_p] \tag{30.206}$$

The state of this system is $x = \begin{bmatrix} x_i \\ x_p \end{bmatrix}$ where x_i is the integrator state and x_p is the plant state. The singular values for the augmented system P_d exhibit a slope of –20 dB/dec at low frequencies as expected (see Fig. 30.10). The minimum singular value crosses zero dB just above 1 rad/s. The maximum singular value crosses zero dB at about 8 rad/s.

Step 2: Design Target Open Loop Transfer Function Matrix $L_o = G_{KF} = C(sI - A)^{-1}H_f$

Next we design a target open loop transfer function matrix $L_o = G_{KF} = C(sI - A)^{-1}H_f$ that has desirable closed loop properties (e.g., sensitivity singular values, pole locations, stability margins, etc.) at the output. To do this, we use Kalman Filtering ideas. Like LQR loops designed without a cross-state-control-coupling penalty, Kalman Filter loops designed in similar fashion exhibit desirable stability robustness margins (e.g., infinite upward gain margin, at least 6 dB downward gain margin, at least ±60° phase margin). This target loop design is carried out as follows:

- Consider the augmented system shown in Fig. 30.11.

  It will be used to design a target loop transfer function matrix $L_o = G_{KF}$ with desirable closed loop properties at the output. To do so, we begin by forming an augmented system $G_{FOL} = C(sI - A)^{-1}L$ with

$$L = \begin{bmatrix} L_L \\ L_H \end{bmatrix} \tag{30.207}$$

$$L_L = [C_p(-A_p)^{-1}B_p]^{-1} \tag{30.208}$$

$$L_H = (-A_p)^{-1}B_pL_L \tag{30.209}$$

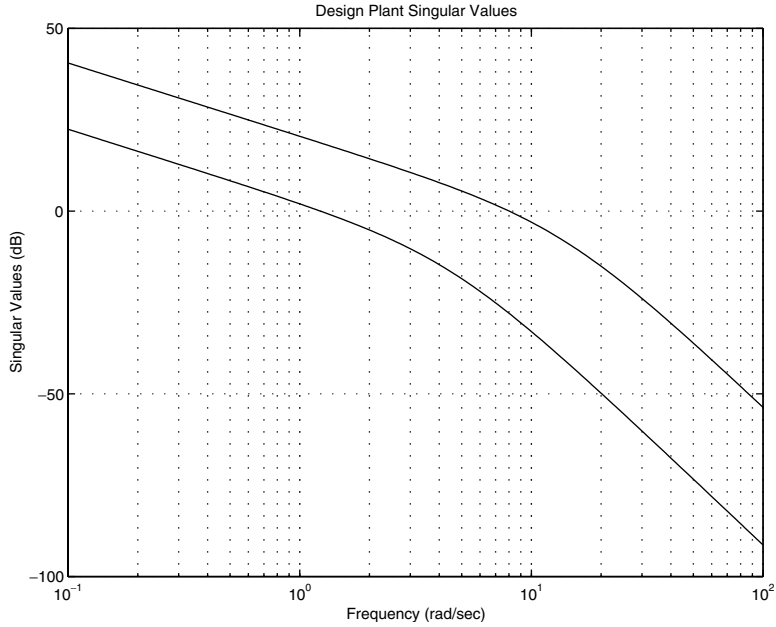



FIGURE 30.10 PUMA 560 robotic manipulator design plant singular values.

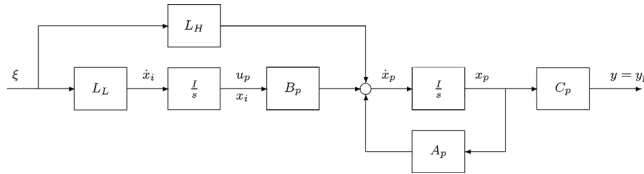



FIGURE 30.11 Augmented system used for designing target loop.

The matrix L_L matches the singular values of $G_{FOL} = C(sI - A)^{-1}L$ at low frequencies. The matrix L_H matches the singular values at high frequencies. Together, L_L and L_H match the singular values of $G_{FOL} = C(sI - A)^{-1}L$ at all frequencies (see Fig. 30.12). Why is this?

This selection for L_L and L_H results in

$$G_{FOL} = C_p(sI - A_p)^{-1}L_H + C_p(sI - A_p)^{-1}B_p\left(\frac{I}{s}\right)L_L \tag{30.210}$$

$$= C_p(sI - A_p)^{-1}\left[L_H + B_p\left(\frac{I}{s}\right)L_L\right] \tag{30.211}$$

$$= C_p(sI - A_p)^{-1}\left[(-A_p)^{-1}B_pL_L + B_p\left(\frac{I}{s}\right)L_L\right] \tag{30.212}$$

$$= C_p(sI - A_p)^{-1}[sI - A_p](-A_p)^{-1}B_pL_L\left(\frac{I}{s}\right) \tag{30.213}$$

$$= C_p(-A_p)^{-1}B_pL_L\left(\frac{I}{s}\right) \tag{30.214}$$

$$= \frac{I}{s} \tag{30.215}$$

The resulting gain crossover frequency in Fig. 30.12 is 1 rad/s, as expected.

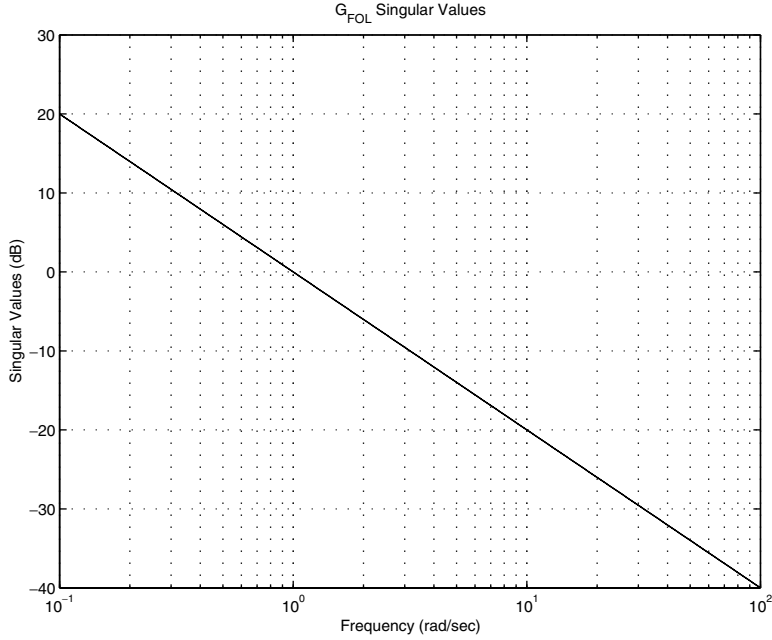



FIGURE 30.12 PUMA 560 robotic manipulator G_{FOL} singular values.

Why match the singular values of G_{FOL} in this manner? From the so-called Kalman Frequency Domain Equality (KFDE), it follows that

$$\sigma_i[I + G_{\mathrm{KF}}(j\omega)] = \sqrt{1 + \frac{1}{\mu}\sigma_i^2[G_{\mathrm{FOL}}(j\omega)]} \tag{30.216}$$

This suggests that by shaping G_{FOL}, we can shape the target $L_o = G_{\mathrm{KF}}$. Specifically, if G_{FOL} is large at low frequencies, then we expect (from KFDE)

$$L_o(j\omega) = G_{\mathrm{KF}}(j\omega) \approx \frac{1}{\sqrt{\mu}} G_{\mathrm{FOL}}(j\omega) \tag{30.217}$$

at low frequencies. This shows that the matrix L should be used for shaping the target loop $L_o = G_{\mathrm{KF}}$ while $\mu > 0$ is used to adjust the target loop bandwidth—decreasing/increasing μ to raise/lower the target loop bandwidth.

Note that through our selection of L, we have made all of the plant's unstable modes uncontrollable through L. Hence, (A, L) is NOT stabilizable! While this might appear to be troublesome, it is not. What matters is that the associated Hamiltonian belongs to *dom*(*Ric*) so that a stabilizing H_f exists. A necessary and suficient condition for this, however, is that (A, C) be detectable and (A, L) has no unobservable modes on the imaginary axis. Since each of these conditions are indeed satisfied, we can use the "are" command to find a stabilizing solution to the FARE.

- Next we solved the FARE with $\Theta = \mu I_{2\times 2}\,(\mu = 0.1)$:

$$AY + YA^T + LL^T - YC^T\Theta^{-1}CY = 0 \tag{30.218}$$

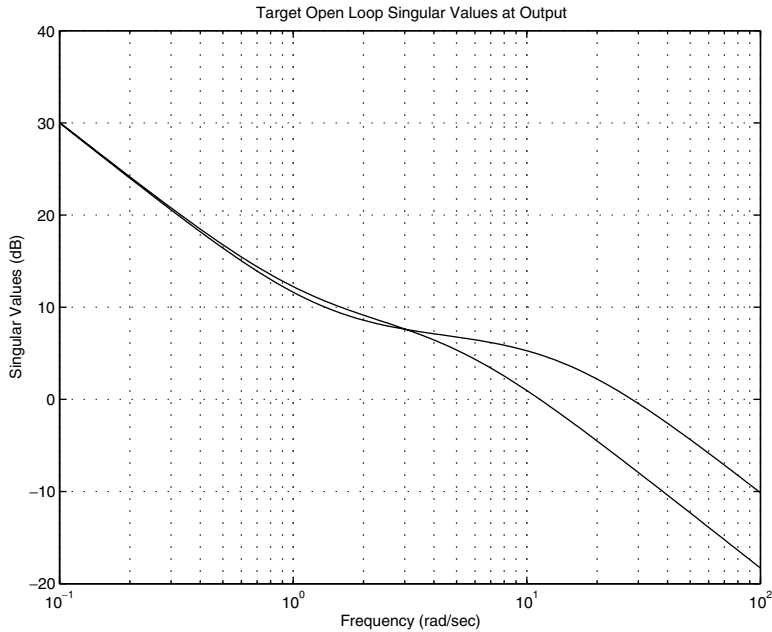



FIGURE 30.13 PUMA 560 robotic manipulator target loop G_{KF} singular values.

for $Y \geq 0$. The "are" command was used to do this, as it returns a stabilizing solution (provided that one exists). We then formed the filter gain matrix

$$H_f = YC^T\Theta^{-1} \tag{30.219}$$

$$= \begin{bmatrix} 2.3635 & 0.0384 \\ 0.4085 & 0.3091 \\ 13.1371 & -4.2300 \\ -4.2300 & 30.4572 \\ 90.2377 & -83.4384 \\ -100.9668 & 467.7679 \end{bmatrix} \tag{30.220}$$

Doing so results in the following target closed loop poles ($\lambda_i(A - H_fC)$):

$$s = -3.1623, -3.1623, -4.5299, -4.5299, -14.1050, -14.1050 \tag{30.221}$$

The singular values for the resulting target open loop transfer function matrix $L_o = G_{KF} = C(sI - A)^{-1}H_f$ are shown in Fig. 30.13.

The target open loop singular values—as expected from the KFDE—are matched at low frequencies with a slope of –20 dB/dec. They remain matched til about 1 rad/s, then they separate. This is expected since $G_{FOL} = I/s$ is not an achievable loop. (Not if closed loop stability matters!) The resulting filter gain matrix provides the necessary bandwidth to stabilize the unstable robotic manipulator, with open loop instabilities at s = 14.1050, 4.5299. One singular value crosses 0 dB just above 10 rad/s, the other just below 30 rad/s. μ was used to adjust the bandwidth.

The corresponding target sensitivity $S_{KF} = [I + G_{KF}]^{-1}$ singular values and complementary sensitivity $T_{KF} = G_{KF}[I + G_{KF}]^{-1}$ singular values are shown in Figs. 30.14 and 30.15, respectively. The associated

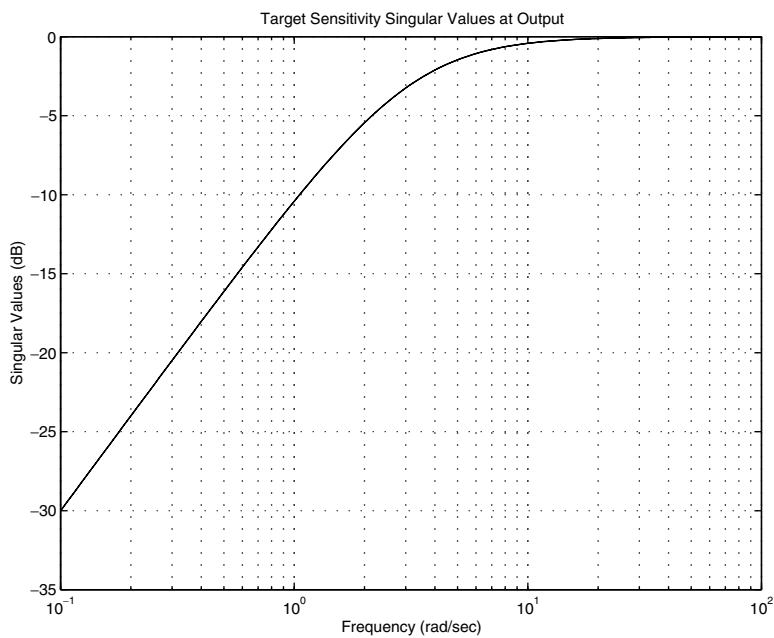



FIGURE 30.14 PUMA 560 robotic manipulator target sensitivity $S_{KF} = [I + G_{KF}]^{-1}$ singular values.

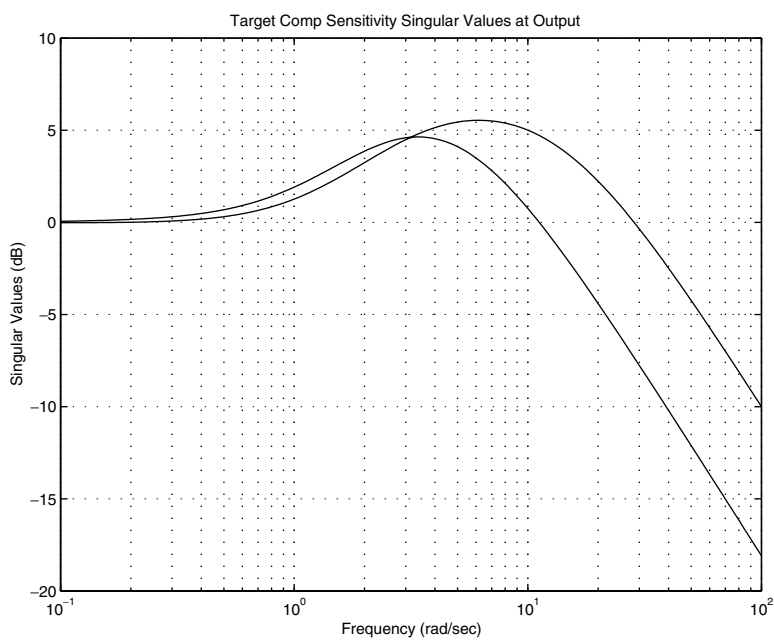



FIGURE 30.15 PUMA 560 robotic manipulator target complementary sensitivity $T_{KF} = G_{KF}[I + G_{KF}]^{-1}$ singular values.

sensitivity and complementary sensitivity singular values are desirable in that they suggest that the target loop will possess:

- good low frequency command following properties,
- good low frequency disturbance attenuation properties,
- good high frequency sensor noise attenuation properties, and
- good MIMO stability margins (nearly infinite upward gain margin, at least 6 dB downward gain margin, and at least ±60° phase margin) at the output.

The complementary sensitivity singular values suggest that a reference command prefilter W would reduce overshoot due to step reference commands. The design of such a filter will be considered below.

Step 3: Solve Cheap Control Problem to Recover Target Loop at Plant Output

Next we solve an appropriately formulated "cheap LQR control problem" that would produce a control gain matrix G_c such that the $\mathcal{H}^2$ optimal model-based compensator $K_d = [A - BG_c - H_f C, H_f, G_c]$ with $P_d = [A, B, C]$ approximates ("recovers") the target loop transfer function matrix $L_o = G_{KF}$; i.e.,

$$P_d K_d \approx L_o = G_{KF} \tag{30.222}$$

This was done by solving the following CARE (using the "lqr" command) with $R = \rho I_{2\times2}(\rho = 10^{-13})$:

$$XA + A^T X + C^T C - XBR^{-1}B^T X = 0 \tag{30.223}$$

for $X \geq 0$ and forming the control gain matrix

$$G_c = R^{-1}B^T X \tag{30.224}$$

$$= \begin{bmatrix} 987.9832 & -543.0034 & 3162945.2928 & 56.9921 & 13941.9005 & 2069.8324 \\ -543.0034 & 3657.5891 & 11.7867 & 3162634.3919 & 2069.7987 & 3765.7978 \end{bmatrix} \tag{30.225}$$

Doing so yields the following closed loop regulator poles ($\lambda_i(A - BG_c)$):

$$s = -440.8808,\ -220.4404 \pm j381.7871,\ -1881.9053,\ -940.9527 \pm j1629.7168 \tag{30.226}$$

All have damping factors greater than or equal to $\zeta = 0.5$. As a practical note to facilitate real-time implementation of the resulting controller, one might use model reduction techniques [10] to remove some of the very high frequency poles in the compensator. Doing so would permit using a larger integration step size in any real-time embedded system or microprocessor implementation.

Step 4: Construct Final Controller K

Next we form the final controller as follows:

$$K = \frac{K_d}{s} \tag{30.227}$$

$$= \frac{[A - BG_c - H_f C, H_f, G_c]}{s} \tag{30.228}$$

$$= [A_k, B_k, C_k] \tag{30.229}$$

A state space representation for this controller is given by

$$A_K = \begin{bmatrix} 0_{2\times2} & G_c \\ 0_{6\times2} & A - BG_c - H_f C \end{bmatrix}, \qquad B_K = \begin{bmatrix} 0_{2\times2} \\ H_f \end{bmatrix} \tag{30.230}$$

$$C_K = \begin{bmatrix} I_{2\times2} & 0_{2\times6} \end{bmatrix} \tag{30.231}$$

With this selection, we have

$$PK = P\frac{K_d}{s} \tag{30.232}$$

$$= P\frac{I_{2\times 2}}{s}K_d \tag{30.233}$$

$$= P_d K_d \tag{30.234}$$

$$\approx L_o = G_{\mathrm{KF}} \tag{30.235}$$

Through this selection of K, we have recovered the target loop transfer function matrix $L_o = G_{\mathrm{KF}}$. That is, K has approximately inverted P (from the right) in order to achieve $PK \approx L_o = G_{\mathrm{KF}}$. An examination of the singular values for the actual loop PK shows that the actual singular values agree with the target singular values up to and beyond 100 rad/s.

Loop Transfer Recovery. Why were we able to recover the target loop? The recovery was permitted by the model-based structure of the compensator K, the Riccati equations used to obtain the gain matrices G_c and H_f, and the fact that the plant $P = [A_p, B_p, C_p]$ (and hence the design plant $P_d = [A, B, C] = P(I/s)$ is minimum phase. The minimum phase condition, specifically, is a suficient condition which guarantees that there exists an orthonormal matrix $U(U^T U = UU^T = I)$ such that

$$\lim_{\rho \to 0^+} \sqrt{\rho} G_c = UC \tag{30.236}$$

This limiting behavior relating the control gain matrix and the design plant's C matrix, however, can be used to prove that loop transfer recovery takes place; i.e.,

$$\lim_{\rho \to 0^+} P_d K_d = \lim_{\rho \to 0^+} PK = L_o = G_{\mathrm{KF}} \tag{30.237}$$

Step 5: Design Command Pre-filter W

The MATLAB command

$$t_0(a - b * g - h * c, h, g) \tag{30.238}$$

can be used to find the compensator's transmission zeros. These are also zeros of the closed loop transfer function matrix from r to y. The final compensator (as well as the target loop G_{KF}) has zeros near $s \approx -1.2$. Given this, a reference command prefilter

$$W = \frac{1.2}{s + 1.2} I_{2\times 2} \tag{30.239}$$

was added outside the loop to filter reference commands. By so doing, we ensure that step reference commands for θ_1 and θ_2 are followed in the steady state (due to integrators in controller) without excessive overshoot during the transient.

Sensitivity Frequency Response

The resulting sensitivity singular values are plotted in Fig. 30.16. The plot suggests that low frequency reference commands r will be followed and low frequency output disturbances d_o will be attenuated. More precisely, reference commands r with frequency content below 0.3 rad/s should be followed to within about 20 dB; that is, with a steady-state error of about 10%. Similarly, output disturbances d_o with frequency content below 0.3 rad/s should be attenuated by approximately 20 dB.

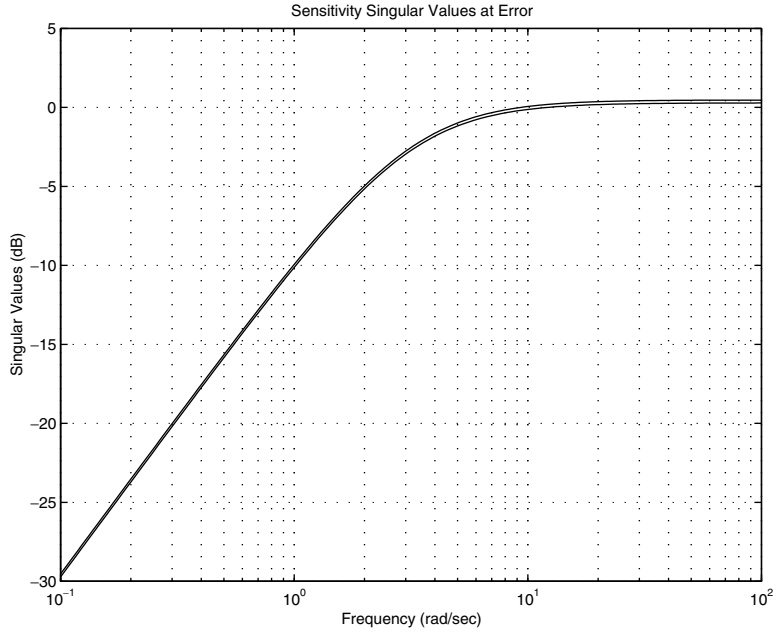



FIGURE 30.16 PUMA 560 sensitivity frequency response at error.

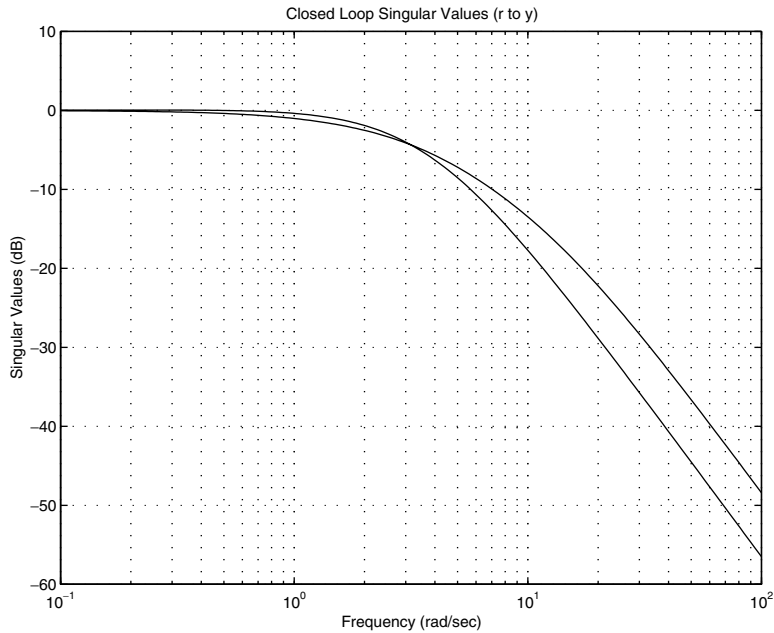



FIGURE 30.17 PUMA 560 reference to output frequency response.

Reference to Output Frequency Response

The transfer function matrix from reference commands r to link angles y is

$$T_{ry} = [I + PK]^{-1} PKW \tag{30.240}$$

Its singular values are plotted in Fig. 30.17. The plot suggests that low frequency reference commands will be followed in the steady state and that little overshoot will result during the transient.

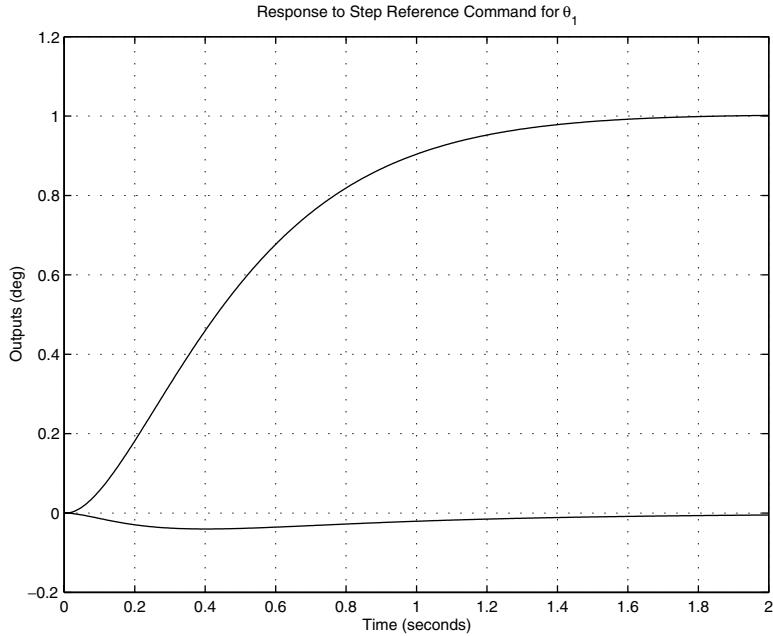



FIGURE 30.18 PUMA 560 outputs: response to θ_1 reference command.

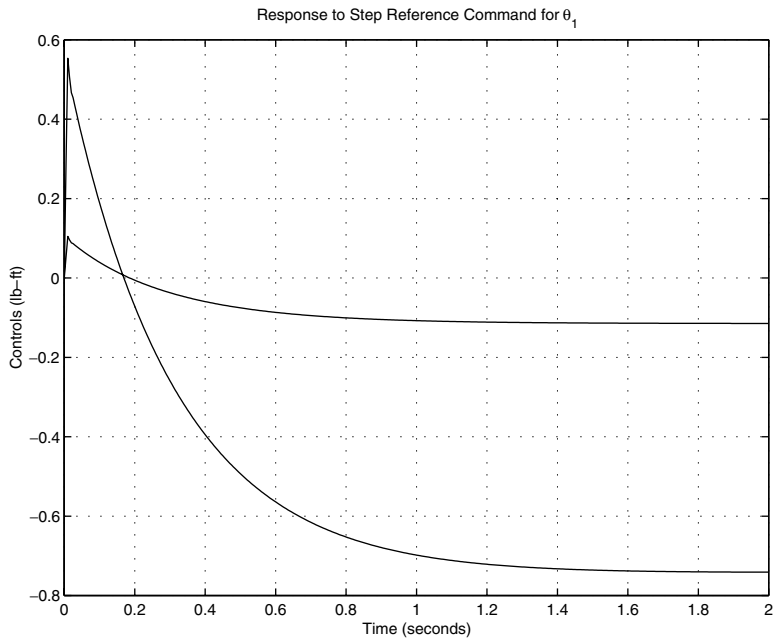



FIGURE 30.19 PUMA 560 controls: response to θ_1 reference command.

Response to Θ_1 Step Reference Command

The response to a unit step θ_1 command is plotted in Fig. 30.18. As expected, θ_1 follows the step command well, with no overshoot and settling in about 1.6 s. The associated θ_2 response is small, indicating little cross coupling in the final closed loop system. The corresponding controls are plotted in Fig. 30.19. They are acceptable in size.

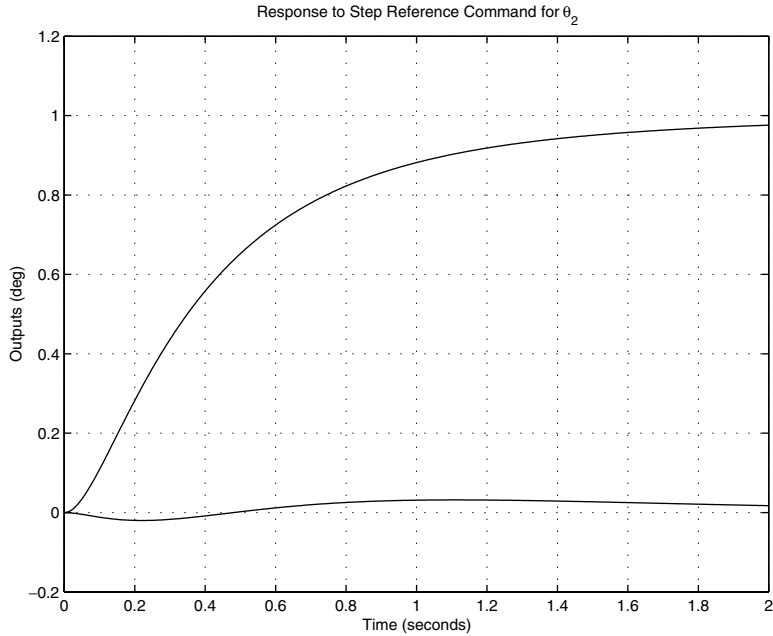



FIGURE 30.20 PUMA 560 outputs: response to θ_2 reference command.

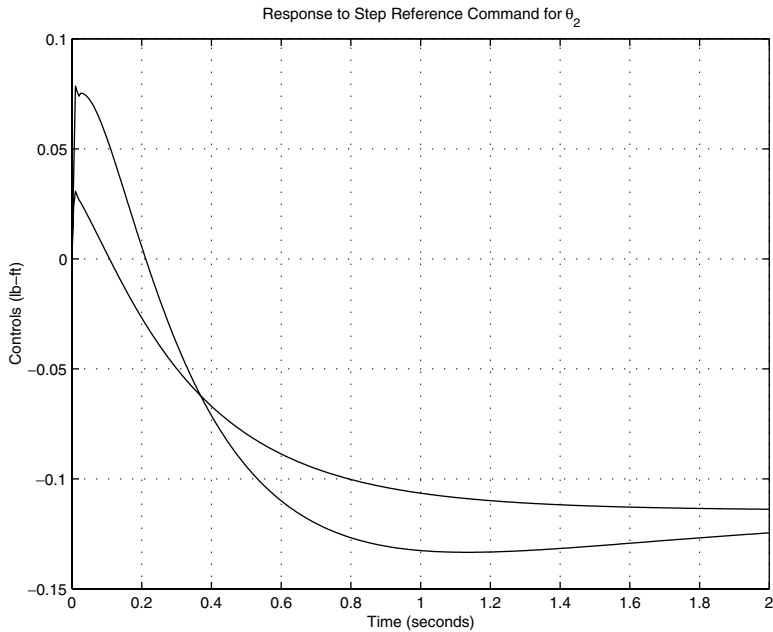



FIGURE 30.21 PUMA 560 controls: response to θ_2 reference command.

Response to Θ_2 Step Reference Command

The response to a unit step θ_2 command is plotted in Fig. 30.20. As expected, θ_2 follows the step command well, with no overshoot and settling in about 3 s. The associated θ_1 response is small, indicating little cross coupling in the final closed loop system. The corresponding controls are plotted in Fig. 30.21. They are acceptable in size. ■

30.4 $\mathcal{H}^2$ State Feedback Problem

This section shows that the methods presented for output feedback may be readily adopted to permit the design of $\mathcal{H}^2$ optimal constant gain state feedback control laws (control gain matrices G_c) as well.

Generalized Plant Structure for State Feedback

For this case, the generalized plant G (including plant P and weighting functions) takes the following form:

$$G = \left[\begin{array}{c|c} G_{11} & G_{12} \\ \hline G_{21} & G_{22} \end{array}\right] = \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & 0_{n_z \times n_w} & D_{12} \\ I_{n \times n} & 0_{n_y \times n_w} & 0_{n_y \times n_u} \end{array}\right] = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] \tag{30.241}$$

This implies that the measured signals y are the states x of the generalized plant G. As such, all of the modes of A are observable through $C_2 = I_{n\times n}$.

State Feedback Assumptions

The standard state feedback assumptions are a subset of those required for the output feedback problem formulation. The state feedback assumptions are as follows.

Assumption 30.2 ($\mathcal{H}^2$ State Feedback Problem)

Throughout this section, it will be assumed that

1. *Plant G_{22} Assumption.* (A, B_2) stabilizable.
2. *Nonsingular Control Weighting Assumption.* $R = D_{12}^T D_{12} > 0$ (D_{12} full column rank).
3. *Regulator Assumption.* $\begin{bmatrix} j\omega I - A & -B_2 \\ C_1 & D_{12} \end{bmatrix}$ has full column rank for all ω. ■

It should be noted that if $D_{12}^T C_1 = 0$, then (3) is equivalent to (A, C_1) having no unobservable imaginary modes. If (A, C_1) is detectable, then this is satisfied.

$\mathcal{H}^2$ Optimal State Feedback Control Law

The $\mathcal{H}^2$ optimal controller is given by

$$K_{\text{opt}} = -G_c \tag{30.242}$$

where the control gain matrix $G_c \in \mathcal{R}^{n_u \times n}$ is given by

$$G_c = R^{-1}[B_2^T X + D_{12}^T C_1] \tag{30.243}$$

where $X \geq 0$ is the unique (at least) positive semi-definite solution of the CARE:

$$(A - B_2 R^{-1} D_{12}^T C_1)^T X + X(A - B_2 R^{-1} D_{12}^T C_1) + C_1^T (I - D_{12}^T R^{-1} D_{12}^T) C_1 - X B_2 R^{-1} B_2^T X = 0 \tag{30.244}$$

The closed loop poles that result from the above constant gain state feedback control law are the eigenvalues of $A - B_2 G_c$. The minimum closed loop norm is given by

$$\min_K \|T_{wz}\|_{\mathcal{H}^2} = \sqrt{trace(B_1^T X B_1)} \tag{30.245}$$

State Feedback Loop Shaping
If one selects

$$B_2 = B \tag{30.246}$$

$$C_1 = \begin{bmatrix} M \\ 0_{n_u \times n} \end{bmatrix} \tag{30.247}$$

$$D_{12} = \begin{bmatrix} 0_{n_y \times n_u} \\ \sqrt{\rho} I_{n_u \times n_u} \end{bmatrix} \tag{30.248}$$

$$R = \rho I_{n_u \times n_u} \tag{30.249}$$

then $D_{12}^T C_1 = 0$ and we have

$$G_c = \frac{1}{\rho} B_2^T X \tag{30.250}$$

where $X \geq 0$ is the unique (at least) positive semi-definite solution of the CARE:

$$A^T X + XA + M^T M - XB\frac{1}{\rho}B^T X = 0 \tag{30.251}$$

The following LQFDE may be derived from the CARE:

$$[I + G_{\text{LQ}}(j\omega)]^H [I + G_{\text{LQ}}(j\omega)] = I + \left[\frac{1}{\sqrt{\rho}} G_{\text{OL}}(j\omega)\right]^H \left[\frac{1}{\sqrt{\rho}} G_{\text{OL}}(j\omega)\right] \tag{30.252}$$

where

$$G_{\text{OL}} = M(sI - A)^{-1} B \tag{30.253}$$

$$G_{\text{LQ}} = G_c(sI - A)^{-1} B \tag{30.254}$$

Given this, the loop shaping ideas discussed earlier are applicable. A designer may use the matrix M and the scalar $\rho > 0$ to shape G_{OL} in an effort to get a desirable loop G_{LQ}. The matrix M, specifically, may be used to match singular values at low frequencies, high frequencies, all frequencies, etc. Assuming that (A, B) is stabilizable and (A, M) has no imaginary modes that are unobservable, a stabilizing solution is guaranteed to exist. Moreover, the resulting G_{LQ} loop will possess nominal sensitivity and stability robustness properties—a consequence of the LQFDE. The resulting control gain matrix G_c may be used within a state feedback loop, a modified state feedback loop, or within a model-based compensator.

30.5 $\mathcal{H}^2$ Output Injection Problem

This section shows how the methods presented for output feedback may be readily adopted to permit the design of $\mathcal{H}^2$ optimal state estimators (filter gain matrices H_f) as well.

Generalized Plant Structure for Output Injection

For this case (dual to the state feedback case), the generalized plant G (including plant P and weighting functions) takes the following form:

$$G = \left[\begin{array}{c|c} G_{11} & G_{12} \\ \hline G_{21} & G_{22} \end{array}\right] = \left[\begin{array}{c|cc} A & B_1 & I_{n\times n} \\ \hline C_1 & 0_{n_z\times n_w} & 0_{n_z\times n_u} \\ C_2 & D_2 & 0_{n_y\times n_u} \end{array}\right] = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] \tag{30.255}$$

This implies that the control signals u directly impact all of the generalized plant states x. As such, all of the modes of A are controllable through $B_2 = I_{n\times n}$.

Output Injection Assumptions

The standard output injection assumptions are a subset of those required for the output feedback problem formulation. The output injection assumptions are as follows.

Assumption 30.3 ($\mathcal{H}^2$ Output Injection Problem)

Throughout this section, it will be assumed that

1. *Plant G_{22} Assumption.* (A, C_2) detectable.
2. *Nonsingular Measurement Weighting Assumption.* $\Theta = D_{21}D_{21}^T > 0$ (D_{21} full row rank).
3. *Filter Assumption.* $\begin{bmatrix} j\omega I - A & -B \\ C_1 & D_{21} \end{bmatrix}$ has full row rank for all ω. ■

It should be noted that if $B_1D_{21}^T = 0$, then (3) is equivalent to (A, B_1) having no uncontrollable imaginary modes. If (A, B_1) is stabilizable, then this is satisfied.

$\mathcal{H}^2$ Optimal Output Injection Law

The $\mathcal{H}^2$ optimal controller is then given by

$$K_{\text{opt}} = -H_f \tag{30.256}$$

where the filter gain matrix $H_f \in \mathcal{R}^{n\times n_y}$ is given by

$$H_f = [YC_2^T + B_1D_{21}^T]\Theta^{-1} \tag{30.257}$$

where $Y \geq 0$ is the unique (at least) positive semi-definite solution of the FARE:

$$(A - B_1D_{21}^T\Theta^{-1}C_2)Y + Y(A - B_1D_{21}^T\Theta^{-1}C_2)^T + B_1(I - D_{21}^T\Theta^{-1}D_{21})B_1^T - YC_2^T\Theta^{-1}C_2Y = 0 \tag{30.258}$$

The closed loop poles that result from the above output injection law are the eigenvalues of $A - H_fC_2$. The minimum closed loop norm is given by

$$\min_K \|T_{wz}\|_{\mathcal{H}^2} = \sqrt{trace\left(C_1YC_1^T\right)} \tag{30.259}$$

where Y is the solution to the FARE.

Estimator (Filter) Loop Shaping

If one selects

$$B_1 = [L \quad 0_{n \times n_y}] \tag{30.260}$$

$$D_{21} = [0_{n_y \times n_u} \quad \sqrt{\mu} I_{n_y \times n_y}] \tag{30.261}$$

$$C_2 = C \tag{30.262}$$

$$\Theta = \mu I_{n_y \times n_y} \tag{30.263}$$

then $B_1 D_{21}^T = 0$ and we have

$$H_f = Y C_2^T \frac{1}{\mu} \tag{30.264}$$

where $Y \geq 0$ is the unique (at least) positive semi-definite solution of the FARE:

$$YA^T + AY + LL^T - YC^T \Theta^{-1} CY = 0 \tag{30.265}$$

Given this, the following KFDE may be derived from the FARE:

$$[I + G_{\mathrm{KF}}(j\omega)][I + G_{\mathrm{KF}}(j\omega)]^H = I + \left[\frac{1}{\sqrt{\mu}} G_{\mathrm{FOL}}(j\omega)\right]\left[\frac{1}{\sqrt{\mu}} G_{\mathrm{FOL}}(j\omega)\right]^H \tag{30.266}$$

where

$$G_{\mathrm{FOL}} = C(sI - A)^{-1} L \tag{30.267}$$

$$G_{\mathrm{KF}} = C(sI - A)^{-1} H_f \tag{30.268}$$

Given this, the loop shaping ideas discussed earlier are applicable. A designer may use the matrix L and the scalar $\mu > 0$ to shape G_{FOL} in an effort to get a desirable loop G_{KF}. The matrix L, specifically, may be used to match singular values at low frequencies, high frequencies, all frequencies, etc. Assuming that (A, C) is detectable and (A, L) has no imaginary modes that are uncontrollable, then a stabilizing solution is guaranteed to exist. Moreover, the resulting G_{KF} loop will possess nominal sensitivity and stability robustness properties—a consequence of the KFDE. The resulting filter (output injection) gain matrix H_f may be used within an estimator (feedback) loop, a modified estimator (feedback) loop, or within a model-based compensator.

30.6 Summary

This chapter has presented a general framework for control system design via $\mathcal{H}^2$ optimization. While the focus has been on continuous time LTI systems, the methods are very flexible and have wide application. They may be used to design constant gain state feedback control laws, constant gain state estimators, dynamic output feedback controllers, and much more. Weighting functions are easily accommodated within the generalized plant framework presented. Such functions may be used to achieve closed loop design objectives. All of the ideas presented may be extended with subtle (all be it very important)

modifications, to accommodate control system design via $\mathcal{H}^\infty$ optimization. Additional details may be found in [8,11].

The methods presented in this chapter may be extended to discrete time linear shift invariant (LSI) systems. Extensions to sampled data systems are also possible [1].

References

1. Chen, T. and Francis, B., *Optimal Sampled-Data Control Systems,* Springer, London, 1995.
2. Dorf, R.C. and Bishop, R.H., *Modern Control Systems,* Addison Wesley, 8th edition, CA, 1998.
3. Doyle, J.C., "Guaranteed margins for LQG regulators," *IEEE Transactions on Automatic Control,* Vol. AC-23, No. 4, August 1978, pp. 756–757.
4. Doyle, J.C., Glover, K., Khargonekar, P.P., and Francis, B.A., "State-space solutions to standard $\mathcal{H}^2$ and $\mathcal{H}^\infty$ control problems," *IEEE Transactions on Automatic Control,* Vol. AC-34, No. 8, 1989, pp. 831–847. Also see *Proceedings of the 1988 American Control Conference,* Atlanta, Georgia, June, 1988.
5. Kalman, R.E., "A new approach to linear filtering and prediction problems," *ASME Journal of Basic Engineering,* Vol. 85, 1960, pp. 34–45.
6. Kalman, R.E. and Bucy, R.S., "New results in linear filtering and prediction problems," *ASME Journal of Basic Engineering,* 1960, pp. 95–108.
7. Kwakernaak, H. and Sivan, R., *Linear Optimal Control Systems,* Wiley-Interscience, New York, 1972.
8. Rodriguez, A.A., *A Practical Neo-Classical Approach to Feedback Control System Analysis and Design,* Control3D, 2000.
9. Spong, M.W. and Vidyasagar, M., *Robot Dynamics and Control,* John Wiley and Sons, New York, 1989.
10. Zhou, K., Doyle, J.C., and Glover, K., *Robust and Optimal Control,* Prentice-Hall, NJ, 1996.
11. Zhou, K. and Doyle, J.C., Essentials of Robust Control, Prentice-Hall, NJ, 1998.

31

Adaptive and Nonlinear Control Design

Maruthi R. Akella
The University of Texas at Austin

31.1 Introduction

The most important challenge for modern control theory is that it should deliver acceptable performance while dealing with poor models, high nonlinearities, and low-cost sensors under a large number of operating conditions. The difficulties encountered are not peculiar to any single class of systems and they appear in virtually every industrial application. Invariably, these systems contain such a large amount of model and parameter uncertainty that "fixed" controllers can no longer meet the stability and performance requirements. Any reasonable solution for such problems must be a suitable amalgamation between nonlinear control theory, adaptive elements, and information processing. Such are the factors behind the birth and evolution of the field of adaptive control theory, strongly motivated by several practical applications such as chemical process control and design of autopilots for high-performance aircraft, which operate with proven stability over a wide variety of speeds and altitudes.

A commonly accepted definition for an adaptive system is that it is any physical system that is designed from an adaptive standpoint![1] All existing stability and convergence results, in the field of adaptive control theory, hinge on the crucial assumption that the unknown parameters must occur linearly within the plant containing known nonlinearities. Conceptually, the overall process makes the parameter estimates themselves as state variables, thus enlarging the dimension of the state space for the original system. By nature, adaptive control solutions for both linear and nonlinear dynamical systems lead to nonlinear time-varying formulations wherein the estimates of the unknown parameters are updated using input–output data. A parameter adaptation mechanism (typically nonlinear) is used to update the parameters within the control law. Given the nonlinearity due to adaptive feedback, there is the need to ensure that the closed-loop stability is preserved. It is thus an unmistakable fact that the fields of adaptive control and nonlinear system stability are intrinsically related to one another and any new insights gained in one

field would be of potential benefit to the other. Many formalisms in nonlinear stability theory can be employed such as the Lyapunov direct method and passivity-based methods. We will first present some important mathematical and analytical tools for studying the stability of nonlinear dynamical systems.

31.2 Lyapunov Theory for Time-Invariant Systems

The Lyapunov direct method is a commonly adopted and arguably one of the most popular methods for proving closed-loop stability in the adaptive control area. It is not restricted to local system behavior and determines the stability properties of the nonlinear system by considering the time evolution of the system solutions with respect to an "energy-like" scalar function, often known as the Lyapunov function.

Consider any dynamical system represented by the following nonlinear autonomous differential equation

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}), \qquad \boldsymbol{f}(0) = 0 \tag{31.1}$$

Obviously $\boldsymbol{x}(t) = 0$ is a solution. A sufficient condition for the existence and uniqueness of solutions for Eq. (31.1) is that $\boldsymbol{f}(\boldsymbol{x})$ be locally Lipschitz, that is,

$$\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{y})\| \le L\|\boldsymbol{x} - \boldsymbol{y}\| \tag{31.2}$$

for all $\boldsymbol{x}$ and $\boldsymbol{y}$ in a finite neighborhood of the origin. We are interested in the stability of the solutions of Eq. (31.1) in the presence of perturbations. Before discussing the main Lyapunov stability theorems, we present some important definitions.

Definition: **Lyapunov stability**
The solution $\boldsymbol{x}(t) = 0$ of Eq. (31.1) is called *stable* in the sense of Lyapunov if for all $\epsilon > 0$, there exists a $\delta(\epsilon) > 0$ such that for all initial conditions satisfying $\|\boldsymbol{x}(0)\| < \delta$, we have $\|\boldsymbol{x}(0)\| < \epsilon$ for $t \in [0, \infty)$. The solution is *unstable* if it is not stable. The solution is *asymptotically stable* if it is stable and there exists a $\delta > 0$ such that every initial condition that satisfies $\|\boldsymbol{x}(0)\| < \delta$ has the property

$$\lim_{t \to \infty} \|\boldsymbol{x}(t)\| = 0$$

The solution is *globally asymptotically stable* if it is asymptotically stable for all initial conditions. These definitions refer to stability of particular solutions of Eq. (31.1) with respect to initial conditions and not to the stability of differential equations.

Definition: **Positive definite and semidefinite functions**
Any continuously differentiable function $V: \mathcal{R}^n \to \mathcal{R}$ is called *positive definite* if (i) $V(0) = 0$ and (ii) $V(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \neq 0$. A function is *positive semidefinite* if condition (ii) is replaced by $V(\boldsymbol{x}) \ge 0$ for all $\boldsymbol{x} \neq 0$.

Theorem: **Lyapunov's stability theorem for time-invariant systems**
If there exists a positive definite function $V: \mathcal{R}^n \to \mathcal{R}$ such that the time derivative of V along the solution of $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$ given by

$$\frac{d}{dt}V = \left[\frac{\partial V}{\partial \boldsymbol{x}}\right]^T \dot{\boldsymbol{x}} = \left[\frac{\partial V}{\partial \boldsymbol{x}}\right]^T \boldsymbol{f}(\boldsymbol{x})$$

is negative semidefinite, then the solution $\boldsymbol{x}(t) = 0$ of Eq. (31.1) is stable. In this case, the solution converges to the set $\{\boldsymbol{x} \in \mathcal{R}^n: \dot{V}(\boldsymbol{x}) = 0\}$. If $\dot{V}$ is negative definite, then the solution is asymptotically stable. Furthermore, if $\dot{V}$ is negative definite and $V(\boldsymbol{x}) \to \infty$ as $\|\boldsymbol{x}\| \to \infty$, then the solution is globally asymptotically stable. The function $V(\boldsymbol{x})$ is called a Lyapunov function for the system described by Eq. (31.1).

Remark: Lyapunov's theorem, though simple to state, has powerful applications in the stability analysis of nonlinear systems. However, since the theorem provides only a sufficient condition in terms of the Lyapunov function, we are often encountered with the difficult problem of finding a suitable Lyapunov function. In the special case when Eq. (31.1) is a stable linear system,

$$\dot{\boldsymbol{x}} = A_m \boldsymbol{x}$$

a quadratic Lyapunov function $V = \boldsymbol{x}^T P \boldsymbol{x}$ exists where P is a symmetric positive definite matrix satisfying the so-called Lyapunov equation

$$A_m^T P + P A_m = -Q \tag{31.3}$$

for any symmetric positive definite Q matrix. On the other hand, there is no general recipe for construction of Lyapunov functions for nonlinear systems. As a rule of thumb, in the case of mechanical systems, "energy-like"quantities are good candidates for a first attempt.

31.3 Lyapunov Theory for Time-Varying Systems

We are now ready to consider the stability of solutions for a time-varying (nonautonomous) differential equation

$$\dot{\boldsymbol{x}} = \boldsymbol{g}(\boldsymbol{x},t), \quad \boldsymbol{g}(0,t) = 0 \quad \forall t \leq 0 \tag{31.4}$$

The function $\boldsymbol{g}$ is assumed to be piecewise continuous with respect to t and locally Lipschitz in $\boldsymbol{x}$ about a neighborhood of the solution $\boldsymbol{x}(t) = 0$. This would guarantee that the origin is an equilibrium for Eq. (31.4). In order to investigate the stability of equilibrium for this nonautonomous system, it is important to recognize that any solution of Eq.(31.4) depends not only on time t but also on the initial time t_0. Thus, we need to revisit our previous definitions of stability.

Definition: **Uniform Lyapunov stability**
The solution $\boldsymbol{x}(t) = 0$ for Eq.(31.4) is *uniformly stable* if for every $\epsilon > 0$, there exists a $\delta(\epsilon) > 0$ that is independent of the initial time t_0 such that

$$\|\boldsymbol{x}(t_0)\| < \delta \quad \text{implies} \quad \|\boldsymbol{x}(t)\| < \epsilon \quad \forall t \geq t_0 \geq 0$$

The solution is *uniformly asymptotically stable* if it is uniformly stable and there is a positive constant ρ independent of t_0 such that $\|\boldsymbol{x}(t)\| \to 0$ as $t \to \infty$ for all $\|\boldsymbol{x}(t_0)\| < \rho$. The solution is *globally uniformly asymptotically stable* if it is uniformly asymptotically stable for all initial conditions.

The main stability theorem for nonautonomous systems requires the definition of certain *class K functions.*

Definition: **Class K functions**
A continuous function α: $[0, a) \to [0, \infty)$ is said to belong to *class K* if it is strictly increasing and $\alpha(0) = 0$. It is said to belong to class K_∞, or radially unbounded, if $a = \infty$ in such a way that $\alpha(r) \to \infty$ as $r \to \infty$.

Theorem: **Lyapunov's stability theorem for time-varying systems**
Consider a set $D = \{\boldsymbol{x} \in \mathcal{R}^n: \|\boldsymbol{x}\| \leq R\}$ about the equilibrium $\boldsymbol{x}(t) = 0$ for Eq. (31.4). If there exists a scalar function V: $\mathcal{R}^n \times \mathcal{R}^+ \to \mathcal{R}$ with continuous partial derivatives such that

(i) $\alpha_1(\|\boldsymbol{x}\|) \leq V(\boldsymbol{x}, t) \leq \alpha_2(\|\boldsymbol{x}\|)$ *positive definite* and *decrescent*

(ii) $\dot{V} = \dfrac{\partial V}{\partial t} + \left[\dfrac{\partial V}{\partial \boldsymbol{x}}\right]^T \boldsymbol{g}(\boldsymbol{x}, t) \leq \alpha_3(\|\boldsymbol{x}\|)$

for all $t \geq 0$, where α_1, α_2, and α_3 are class K functions, then the equilibrium point $\boldsymbol{x} = 0$ is uniformly asymptotically stable.

Remark: Note that in order to show stability for nonautonomous systems, it is necessary to bound the function $V(\boldsymbol{x}, t)$ by the class K functions that do not depend upon time t. A detailed treatment of all the definitions and proof for this theorem can be found in Slotine and Li[2] and Khalil.[3]

Remark: In the recent years, several interesting converse Lyapunov results have been obtained. In particular, for every uniformly stable (or uniformly asymptotic stable) system, there exists a positive definite Lyapunov function with a negative semidefinite time derivative (see Sastry and Bodson[4]). These results are particularly useful from a closed-loop performance point of view because they allow us to explicitly estimate the convergence rates in some cases of nonlinear adaptive control systems.

The application of Lyapunov's stability theorem for nonautonomous systems arising out of adaptive control often leads us to negative semidefinite time derivatives of the Lyapunov function. Therefore, asymptotic stability analysis is a much harder problem and the following result, known as Barbalat's Lemma, is extremely useful in such situations.

Lemma: **Barbalat.**
Consider a uniformly continuous function $\phi: \mathcal{R} \rightarrow \mathcal{R}$ defined at all real values of $t \geq 0$. If

$$\lim_{t\rightarrow\infty} \int_0^t \phi(s)\, ds$$

exists and is finite, then $\phi(t) \rightarrow 0$ as $t \rightarrow \infty$.

Remark: A consequence of this result is that if $\phi \in \mathcal{L}_2$ and $\phi \in \mathcal{L}_\infty$, then $\phi(t) \rightarrow 0$ as $t \rightarrow \infty$ (see Slotine and Li[2] and Tao[5] for discussion and proof).

31.4 Adaptive Control Theory

In contrast to a fixed or ordinary controller, an adaptive controller is one with adjustable parameters and an adjustment mechanism. The following are some basic concepts that are necessary for any discussion of adaptive control theory.

Regulation and Tracking Problems

The desired objective for any control problem is to maintain the plant output either at its desired value or within specified/acceptable bounds of the desired value. If these desired values are constant with respect to time, we have a regulation problem, otherwise it is a tracking problem.

Certainty Equivalence Principle

This principle has been the bedrock of most adaptive control design methods and has received considerable attention during the past two decades.[4,6,7] Adaptive controllers based on this approach are obtained by independently designing a control law that meets the control objective assuming complete knowledge of all the unknown plant parameters (deterministic case), along with a parameter update law, which is usually a differential equation that generates online parameter estimates that are used to replace the unknown parameters within the control law. Such a controller would have perfect output tracking capability in the case when the plant parameters are exactly known. In the presence of parameter uncertainty, the adaptation mechanism will adjust the controller parameters so that the tracking objective is asymptotically achieved. The main issue, thus, in adaptive controller design is to synthesize the adaptation mechanism (parameter update law) that will guarantee that the control system remains stable and the output tracking error converges to zero as the parameter values are updated.

Direct and Indirect Adaptive Control

There exist two philosophically distinct approaches within adaptive control for plants containing unknown or uncertain parameters. The first of those is the so-called *direct approach* where the controller parameters are directly adjusted by the adaptation mechanism in such a way to optimize some pre-specified performance index based on the output. The second approach is the *indirect approach* wherein plant parameters are directly estimated and updated by the adaptation law and these estimated values are then used to compute the controller parameters. Direct adaptive control eliminates the need for this additional computation. Consequently, indirect adaptive control is plant parameter adaptive, whereas direct adaptive control is output performance adaptive. The plant parameter identification process is explicit within the direct approach while implicit in the direct approach. Hence, they have also been referred to as explicit and implicit approaches. In both of these cases, the controller structure remains the same and is determined from the certainty equivalence principle.

Model Reference Adaptive Control (MRAC)

The MRAC framework consists of four parts: (i) the plant containing the unknown parameters, (ii) a suitable reference model for specifying the desired output characteristics, (iii) a feedback control law that contains adjustable parameters, and (iv) an adaptation mechanism that updates the adjustable parameters within the control law. A schematic sketch for this framework is shown in Fig. (31.1).

The plant is assumed to have a known structure with unknown parameters. For the case of linear systems, this means that the number of poles and zeros are assumed to be known, but the exact locations of poles and zeros are unknown. In the case of nonlinear systems, the structure of the governing equations of motion is assumed to be known, but some of the parameters appearing linearly within those equations can be unknown. The reference model specifies the desired output behavior expected from the plant as a result of the external reference input. It provides the ideal plant response which the adaptation mechanism should seek to track while updating the parameter estimates. The choice of the reference plant lies at the heart of any MRAC design and any acceptable selection must essentially satisfy two crucial requirements. The first of these requirements is that the reference model must accurately reflect the closed-loop performance specifications such as rise time, settling time, overshoot, and other transient performance characteristics. The other requirement is that given the assumed structure of the plant dynamics, the reference model's output behavior should be asymptotically achievable by the adaptive control system implying certain extra conditions on the relative degree of the reference model and persistent excitation conditions on the reference input. The controller structure is dictated by the certainty equivalence

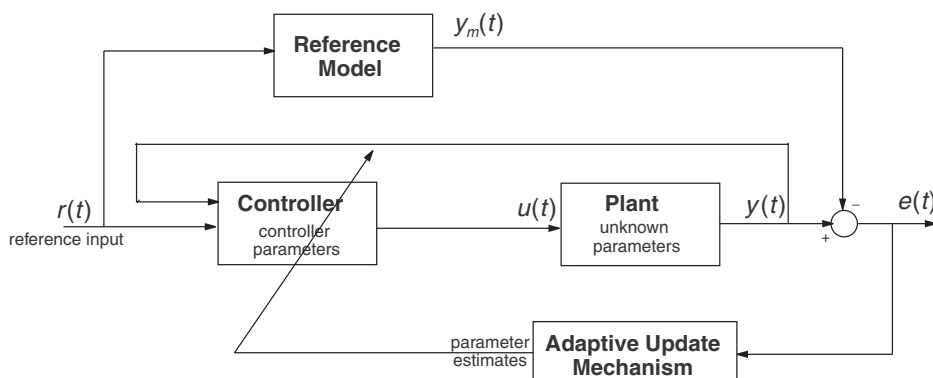



FIGURE 31.1 The model reference adaptive control framework.

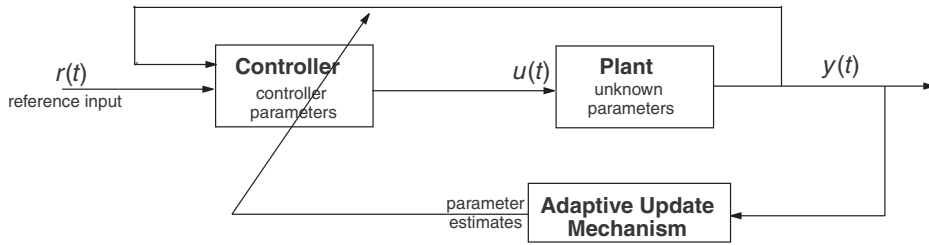



FIGURE 31.2 The self-tuning control architecture.

principle and both direct and indirect parameter update procedures can be adopted within the MRAC framework. Much of the work in this area deals with continuous time systems.

Self-Tuning Controller (STC)

In contrast to MRAC, there is no reference model in the STC design. A schematic sketch is shown in Fig. 31.2. In this formulation, the controller parameters of the plant parameters are estimated in real time, depending on whether it is a direct or indirect approach. These estimates are then used as if they are equal to the true parameters (certainty equivalence design). Parameter estimation involves finding the *best-fit* set of parameters based on the plant input–output data. This is different from the MRAC parameter adaptation scheme, where the parameter estimates are updated in such a way to achieve asymptotic tracking between the tracking error between the plant and the reference model. In several STC estimation schemes, it is also possible to quantify a measure of the quality of the parameter estimates, which can be used in the design of the controller. Many different combinations of the estimation methods can be adopted and can be applied to both continuous time and discrete time plants. Due to the "separation" between parameter estimation and control in STC, there is greater flexibility in design. However, stability and convergence are difficult to prove and stronger conditions on input signals are required (persistent excitation) to guarantee parameter convergence. Historically speaking, STC designs arose in the study of the stochastic regulation problem and much of the literature is devoted to discrete time plants using an indirect approach. In spite of the seeming difference between MRAC and STC, a direct correspondence exists between problems from both the areas.[8]

31.5 Nonlinear Adaptive Control Systems

For the most general case of nonlinear systems, there exists very limited theory in the field of adaptive control. Even though there is great interest in this area due to potential applications in a wide variety of complex machanical systems, theoretical diffculties exist because of the lack of general analysis tools. However, some important special cases are well understood by now, and we summarize the conditions that these classes of systems satisfy:

1. The unknown parameters within the nonlinear plant are linearly parameterized.
2. The complete state vector is measured.
3. When the unknown parameters are assumed known, the control input can cancel all the nonlinearities in a feedback-linearization sense and any remaining internal dynamics should be stable. The adaptive design is then accomplished by certainty equivalence.

We now show a typical nonlinear MRAC methodology to deal with a situation in which the nonlinear plant model has unknown parameters. Consider the nonlinear system

$$\dot{\boldsymbol{x}} = \theta \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{u} \tag{31.5}$$

where θ is a constant and unknown matrix parameter, and $\boldsymbol{f}$ is a known and differentiable nonlinear vector function. In analogy with the MRAC methodology, we assume that it is desired to have the state $\boldsymbol{x}$ asymptotically track the state $\boldsymbol{x}_m$ of a reference system that satisfies

$$\dot{\boldsymbol{x}}_m = A_m\boldsymbol{x}_m + \boldsymbol{r} \tag{31.6}$$

where $\boldsymbol{r}(t)$ is any piecewise continuous and bounded reference input and A_m is a Hurwitz matrix. Introduce an error vector $\boldsymbol{e} = \boldsymbol{x} - \boldsymbol{x}_m$ so that the error dynamics can be established by taking the difference between Eqs. (31.5) and (31.6) as follows:

$$\dot{\boldsymbol{e}} = \theta\boldsymbol{f}(\boldsymbol{x}) - A_m\boldsymbol{x}_m - \boldsymbol{r} + \boldsymbol{u} \tag{31.7}$$

If the parameter θ is assumed to be known, selecting the control input $\boldsymbol{u} = A_m\boldsymbol{x} + \boldsymbol{r} - \theta\boldsymbol{f}(\boldsymbol{x})$ would render the following structure for the error dynamics:

$$\dot{\boldsymbol{e}} = A_m\boldsymbol{e}$$

which would achieve the control objective. However, such a choice of control law is not impossible because θ is unknown. Hence we retain the same structure for the control law except for replacing θ by its time-varying estimate $\hat{\theta}$ so that the certainty-equivalence-based adaptive control law is given by

$$\boldsymbol{u} = A_m\boldsymbol{x} + \boldsymbol{r} - \hat{\theta}\boldsymbol{f}(\boldsymbol{x}) \tag{31.8}$$

Application of the control law in Eq. (31.7) leads us to the following closed-loop error dynamics:

$$\dot{\boldsymbol{e}} = A_m\boldsymbol{e} - \tilde{\theta}\boldsymbol{f}(\boldsymbol{x}) \tag{31.9}$$

where we have introduced the variable $\tilde{\theta}(t)$ to represent the parameter estimation error $\hat{\theta}(t) - \theta$. There are two things remaining to be done: (i) to show the stability and asymptotic convergence of $\boldsymbol{e}(t)$ to zero as $t \to \infty$, (ii) to provide an appropriate parameter adaptation mechanism for $\hat{\theta}(t)$. We accomplish both these tasks by adopting the Lyapunov method. Given that A_m is Hurwitz, for any choice of symmetric and positive definite matrix Q, there exists a symmetric, positive definite matrix P that satisfies the Lyapunov equation given in Eq. (31.3). Choosing a Lyapunov function in terms of such a P matrix,

$$V = \boldsymbol{e}^T P\boldsymbol{e} + \mathrm{tr}[\tilde{\theta}^T\Gamma^{-1}\tilde{\theta}] \tag{31.10}$$

where Γ is a symmetric positive definite learning rate matrix. Taking the time derivative of V along the solutions of Eq. (31.9) we find that

$$\dot{V} = \boldsymbol{e}^T(PA_m + A_m^T P)\boldsymbol{e} - 2\boldsymbol{e}^T P\tilde{\theta}\boldsymbol{f}(\boldsymbol{x}) + 2\ \mathrm{tr}[\tilde{\theta}^T\Gamma^{-1}\dot{\hat{\theta}}] \tag{31.11}$$

Using several matrix trace identities,[9] it is possible to show that

$$\boldsymbol{e}^T P\tilde{\theta}\boldsymbol{f}(\boldsymbol{x}) = \mathrm{tr}[P\tilde{\theta}\boldsymbol{f}(\boldsymbol{x})\boldsymbol{e}^T] = \mathrm{tr}[\tilde{\theta}\boldsymbol{f}(\boldsymbol{x})\boldsymbol{e}^T P] = \mathrm{tr}[\tilde{\theta}^T P\boldsymbol{e}\boldsymbol{f}^T(\boldsymbol{x})]$$

so that we can combine the last two terms on the right-hand side of Eq. (31.11) as follows:

$$\dot{V} = \boldsymbol{e}^T \underbrace{(PA_m + A_m^T P)}_{-Q} \boldsymbol{e} + 2\ \mathrm{tr}\ [\tilde{\theta}^T\{\Gamma^{-1}\dot{\hat{\theta}} - P\boldsymbol{e}\boldsymbol{f}^T(\boldsymbol{x})\}] \tag{31.12}$$

Since θ is constant, $\dot{\theta} = 0$ and $\dot{\hat{\theta}} = \dot{\tilde{\theta}}$. Thus, if the adaptive law for updating $\hat{\theta}$ is chosen as

$$\dot{\hat{\theta}} = \Gamma P \boldsymbol{e} \boldsymbol{f}^T(\boldsymbol{x}) \tag{31.13}$$

then the derivative of the Lyapunov function in Eq. (31.12) becomes

$$\dot{V} = -\boldsymbol{e}^T Q \boldsymbol{e} \tag{31.14}$$

which is negative semidefinite, but not negative definite. This implies that $V(t) \leq V(0)$ for all $t \geq 0$, and thus, $\boldsymbol{e}$ and $\hat{\theta}$ must be bounded. This further implies that $\boldsymbol{x} = \boldsymbol{e} + \boldsymbol{x}_m$ also is bounded. Also, $V \geq 0$ and $\dot{V} \leq 0$, which means $\lim_{t\to\infty} V(t) \doteq V_\infty$ exists and is finite. Now,

$$\int_0^\infty \dot{V}(\tau)d\tau = -\int_0^\infty \boldsymbol{e}^T(\tau)Q\boldsymbol{e}(\tau)d\tau = V_\infty - V(0)$$

implying that $\boldsymbol{e} \in \mathcal{L}_2 \cap \mathcal{L}_\infty$. From Eq. (31.9), it is obvious that $\dot{\boldsymbol{e}} \in \mathcal{L}_\infty$. Thus, we can invoke Barbalat's lemma to claim that $\boldsymbol{e}(t) \to 0$ as $t \to \infty$. Notice, however, that the parameter error $\tilde{\theta} = \hat{\theta} - \theta$ will not necessarily converge to zero. True parameter convergence can occur only when the reference input $\boldsymbol{r}(t)$ satisfies certain uniform observability and persistent excitation conditions[4].

Remark: Note in the above MRAC design that while stability and tracking error convergence are guaranteed for any value of A_m, Q, and Γ, the performance of the controller will depend critically on the learning rate Γ. "Smaller" learning rates mean that the adaptation will be slow leading to large tracking errors and large transients. Conversely, the upper limit on the learning rate is limited by the presence of unmodeled dynamics, because too large a value for the learning rate will lead to highly oscillatory parameter estimates that can adversely excite the high frequency unmodeled plant dynamics.

Remark: The controller design methodology is based upon three crucial steps: (i) finding the appropriate controller structure in the spirit of feedback linearization, (ii) derivation of the tracking error dynamics that depend upon the parameter error terms, and (iii) finding a suitable Lyapunov function that can be used to derive the parameter update law such that the tracking error will go to zero. Determining the controller structure for the known parameter case is probably the most crucial step within any adaptive design because it turns out that adaptive feedback linearization cannot be alway applied to systems that are linearizable by feedback in the known parameter case. This happens because higher derivatives of the parameter estimates appear in the control law for systems of higher order, making difficult the application of the certainty equivalence principle. Other relatively new approaches deviate from the conventional certainty equivalence principles by adopting integrator backstepping, nonlinear damping, and tuning functions.[10] In these methods, the adaptive law estimates the unknown plant parameters directly, thereby permitting full utilization of any prior knowledge and therefore eliminating the possibility for overparameterization introduced by traditional direct MRAC methods. The design methodology and stability proof are obtained through a recursive process,[10,11] an overview for which can be obtained from Kokotovic[12] and subsequent results by his research group.

Remark: Given the fact that there always exist model errors and other unknown disturbance effects in addition to the unknown parameters, adaptive control solutions would have to address the robustness question. Since the parameter error is always unknown, the Lyapunov function time derivative is always negative semidefinite. This implies that the closed-loop equations are not exponentially stable, nor even uniformly asymptotically stable. Any external or unmodeled disturbance would immediately make $\dot{V}$ indefinite, and most methods that modify the stability proof for robustness are fixed to ensure $\dot{V}$ to be negative outside a compact neighborhood of the equilibrium state. By introducing an additional term in the adaptive law Eq. (31.13), (referred to as σ-modification), Ioannou[7] accomplishes robust stability. This method, though very popular, suffered from the drawback that when the disturbance is absent, the tracking error would not converge to zero. To overcome this problem, other schemes such as the ϵ-modification[6] have been suggested to ensure robustness within the adaptive designs.

31.6 Spacecraft Adaptive Attitude Regulation Example

Consider the problem of a rigid spacecraft with an initial nonzero attitude and body angular velocity vector that has to be brought to rest at a zero attitude vector. This rigid body adaptive attitude regulation problem based on the feedback linearization approach has been derived by Schaub, Akella, and Junkins.[13] The governing equations are described by Euler's rotational equations of motion and the desired linear closed-loop dynamics (LCLD) can be of either PD or PID form.[13, 14] Only a crude estimate of the moment of inertia matrix is assumed to be known. An adaptive control law is presented, which includes an integral feedback term in the desired closed-loop dynamics and achieves asymptotic stability even in the presence of unmodeled external disturbances.

The resulting simulation is illustrated in Fig. 31.3. The attitude vector is specified in terms of the modified Rodrigues parameter (MRP) whose components σ_i are shown in Fig. 31.3(a). Without any adaptation, the open-loop control is still asymptotically stable. However, the transient attitude errors don't match those of the desired LCLD well at all. With adaptation turned on, the performance matches that of the ideal LCLD very closely.

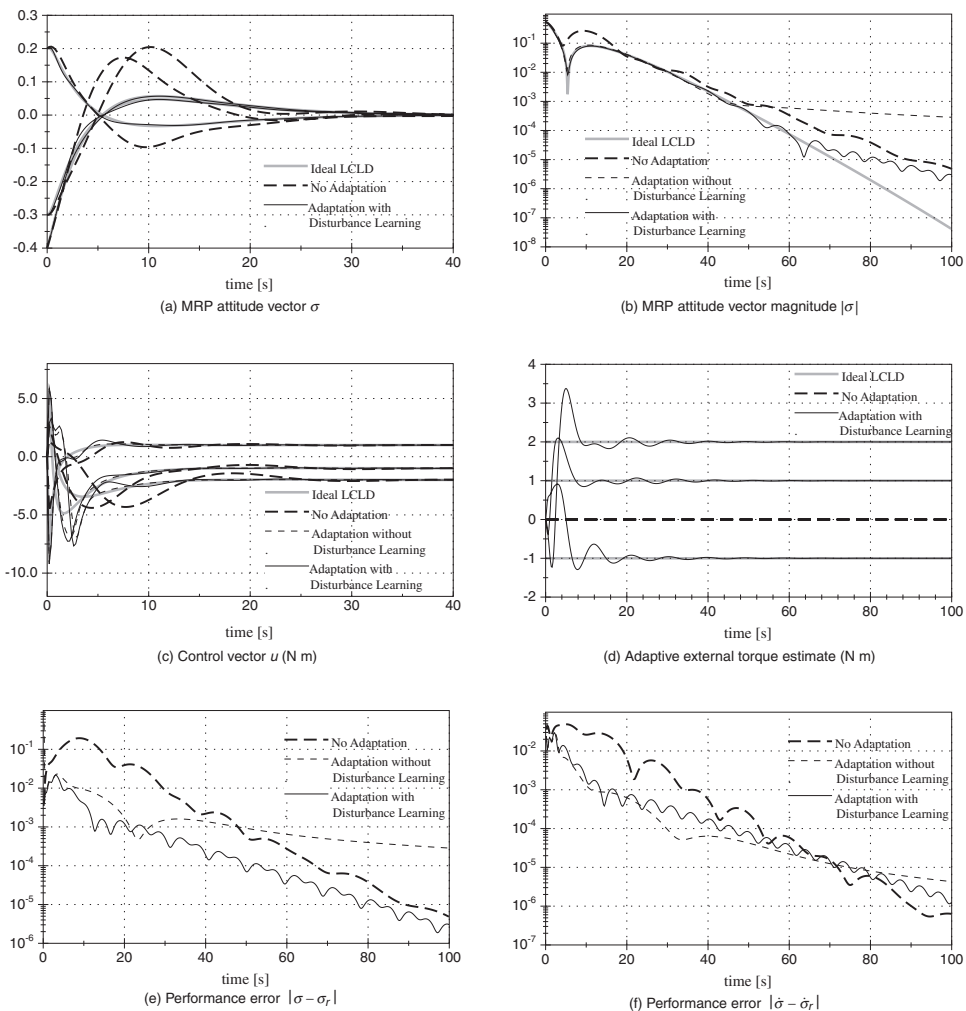



FIGURE 31.3 Rigid body stabilization while enforcing LCLD in the presence of large inertia and external disturbance ignorance.

Figure 31.3(b) shows the magnitude of the MRP attitude error vector σ on a logarithmic scale. Again the large transient errors of the open-loop, adaptation-free control law are visible during the first 20 s of the maneuver along with the good final convergence characteristics. The ideal LCLD performance is indicated again through the dotted line. Two versions of the adaptive control law are compared here, which differ only by whether or not the external disturbance is adaptively estimated too. On this figure both adaptive laws appear to enforce the desired LCLD very well for the first 40 s of the maneuver. After this the adaptive law without disturbance learning starts to decay at a slower rate, slower even than the open-loop (nonadaptive) solution. Including the external disturbance, adaptation clearly improves the final convergence rate. Note, however, that neither adaptive case starts to deviate from the ideal LCLD case until the MRP attitude error magnitude has decayed to roughly 10^{-3}. This corresponds to having a principal rotation error of roughly 0.23°. With external disturbance adaptation, the tracking error at which the LCLD deviations appear is about two orders of magnitude smaller.

The performance of the adaptive control law can be greatly varied by choosing different learning rates. However, since *large* initial inertia matrix and external disturbance model errors are present, the adaptive learning rates were reduced to avoid radical transient torques. The control torque vector components u_i for various cases are shown in Fig. 31.3(c). The open-loop torques don't approach the ideal LCLD torque during the transient part of the maneuver. The torques required by either adaptive case are very similar. The difference is that the case with external disturbance learning is causing some extra oscillation of the control about the LCLD case. However, note that with the chosen adaptive learning rates neither control law exhibits any radical transient torques about the ideal LCLD torque profile. Figure 31.3(d) illustrates that the adaptive external disturbance estimate $\boldsymbol{F}_e$ indeed asymptotically approaches the true external disturbance $\boldsymbol{F}_e^*$. By reducing the external disturbance adaptive learning rate γ_{F_e} the transient adaptive estimate errors are kept within a reasonable range.

Figure 31.3(f) shows the absolute performance error in attitude rates. Both cases with adaptation added show large reductions in attitude rate errors compared to the nonadaptive case.

The purpose of the adaptive control discussed in this example is to enforce the desired LCLD. The previous figures illustrate that the resulting overall system remains asymptotically stable. Figure 31.3(e) illustrates the absolute performance error between the actual motion $\sigma(t)$ and the desired linear reference motion $\sigma_r(t)$. This figure demonstrates again the large performance error that results from using the open-loop control law with the incorrect system model. Adding adaptation improves the transient performance tracking by up to two orders of magnitude. Without including the external disturbance learning, the final performance error decay rate flattens out. This error will decay to zero. However, with the given learning gains, it does so at a slower rate than if no adaptation is taking place. Adding the external disturbance learning greatly improves the final performance error decay since the system is obtaining an accurate model of the actual constant disturbance. If the initial model estimates were more accurate, more aggressive adaptive learning rates could be used, resulting in even better LCLD performance tracking. This simulation illustrates though that even in the presence of large system uncertainty it is possible to track the desired LCLD very well.

31.7 Output Feedback Adaptive Control

In contrast to the state-space approaches, the input–output approach treats the plant as a black box that transforms the applied inputs into the corresponding output space. Stability theory for nonlinear systems from an input–output viewpoint is important in the context of adaptive output feedback control design. Solution to the problem of adaptive observer design involving state estimation of systems with unknown parameters is often the stepping stone towards resolving the output feedback control problem. There has been fairly recent breakthroughs in this area where the nonlinear adaptive observer design procedure has been extended to a slightly more general case of systems where the coefficients of the unknown parameters can depend on the entire state, and not just on the measured part.[15]

To a large extent, some powerful results have been made possible by exploiting certain "passivity-like"conditions coupled with the usual persistent excitation conditions. Crucial to this discussion is the

concept of passivity, which is really an abstract representation of the idea of energy dissipation in both linear and nonlinear systems. Passive systems are most common in mechanical and electrical engineering applications. A mechanical system consisting of masses, springs, and viscous dashpots is a common example for a passive system. We now give the following definitions.

Definition: **Truncation of a signal**
Let Y be the space of real-valued functions defined on $[0, \infty)$. Let $\boldsymbol{x}$ be an element of Y. Then the *truncation* of $\boldsymbol{x}$ at some $T > 0$ is defined by

$$\boldsymbol{x}_T(t) = \begin{cases} \boldsymbol{x}(t) & \text{for } 0 \le t \le T \\ 0 & \text{for } t > T \end{cases}$$

Definition: **Extended space**
If X is a normed linear subspace of Y, then the *extended space* X_e is defined by the set

$$\{\boldsymbol{x} \in Y : \boldsymbol{x}_T \in X \text{ for some fixed } T \ge 0\}$$

The extended $\mathcal{L}_2$ space is denoted by $\mathcal{L}_{2e}$.

Definition: **Scalar product between two signals**
The scalar product between two real-valued time signals $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{L}_{2e}$ is defined as

$$\langle \boldsymbol{x} \,|\, \boldsymbol{y} \rangle = \int_0^\infty \boldsymbol{x}^T(\tau)\boldsymbol{y}(\tau)\, d\tau = \int_0^T \boldsymbol{x}^T(\tau)\boldsymbol{y}(\tau)\, d\tau$$

Definition: **Passive systems**
A system with input $\boldsymbol{u}(t)$ and output $\boldsymbol{y}(t)$ is *passive* if

$$\langle \boldsymbol{y} | \boldsymbol{u} \rangle \ge 0$$

The system is *input strictly passive* if $\exists\, \epsilon > 0$ such that

$$\langle \boldsymbol{y} | \boldsymbol{u} \rangle \ge \epsilon \|\boldsymbol{u}\|^2$$

The system is said to be *output strictly passive* if $\exists\, \epsilon > 0$ such that

$$\langle \boldsymbol{y} | \boldsymbol{u} \rangle \ge \epsilon \|\boldsymbol{y}\|^2$$

31.8 Adaptive Observers and Output Feedback Control

We now state the nonlinear adaptive observer problem formulated by Besancon[15]:

$$\begin{aligned} \dot{\boldsymbol{x}} &= \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u}, t) + \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{u}, t)\theta \\ \boldsymbol{y} &= \boldsymbol{h}(\boldsymbol{x}) \end{aligned} \tag{31.15}$$

where functions $\boldsymbol{f}$ and $\boldsymbol{g}$ are C^∞ with respect to all their arguments and θ is a constant and unknown parameter. Variables $\boldsymbol{x}$, $\boldsymbol{u}$, and $\boldsymbol{y}$ respectively denote the state, input, and output vectors. The input signals may be assumed to belong to some set of measurable and bounded functions. By the phrase adaptive observer, we imply the problem of reconstructing a state estimate $\hat{\boldsymbol{x}}(t)$ using the input $\boldsymbol{u}$ and output $\boldsymbol{y}$ in the presence of the unknown parameter θ such that $\lim_{t\to\infty}\|\hat{\boldsymbol{x}}(t) - \boldsymbol{x}(t)\| = 0$. The conditions for the

existence of such an observer are now available,[15] which can be stated as follows. If a corresponding observer exists in the case when θ is known, and if this deterministic case observer is such that when a parameter error $\tilde{\theta} \doteq \hat{\theta} - \theta$ is made and the state estimation error system is passive between the "input" $\tilde{\theta}$ and the output error $\boldsymbol{h}(\hat{\boldsymbol{x}}) - \boldsymbol{y}$, then an asymptotic state observer can be designed even when θ is unknown. In addition to this passivity requirement, parameter error convergence, as usual, would further need persistence of excitation with respect to $\boldsymbol{u}$.

This powerful result finds immediate applications within the problem of spacecraft attitude tracking in the absence of angular velocity measurements.[16] It is now well known that the governing equations of the rigid-body attitude control problem in terms of the MRP vector satisfy certain passivity conditions[17, 18] between the angular velocity vector and the MRP vector. A very important consequence of passivity in this context is the fact that feedback control laws for attitude control can be implemented in a Lyapunov-based construction without requiring angular velocity measurements. In such a case, the only signal needed for feedback purposes would be the attitude vector. The resulting control laws provide *almost* global asymptotic stability in the sense of Tsiotras.[18]

31.9 Concluding Remarks

Historically speaking, the development and application of modern adaptive control theory for generic nonlinear systems adopted the philosophical approach of extending existing linear system methodologies. In some limited cases such as regulator theory, this approach of paralleling linear system methods has been highly successful. However, obtaining the same degree of success has been elusive in other research areas such as trajectory tracking, controller synthesis, and state reconstruction.

It is not difficult to fathom the reason for this bottleneck. Nonlinear systems occur in a vast variety of ways, and not all of them can be handled by simple extensions to existing linear adaptive control methodologies. One promising approach for the purpose of future research would be to specialize the study to mechanical systems, thereby restricting the class of nonlinear systems considered, and thus enabling the introduction of "structure" and additional constraints. Whereas in the case of output feedback control for general nonlinear systems, separate designs of stable observers and controllers do not necessarily guarantee stability for their combination (no separation principle), some structured approaches utilizing state transformations have already been shown to help recover the separation properties in some cases.[15] As a result, these so-called structured approaches also enabled the formulation of global and semi-global tracking controllers based on output (partial state) feedback. It is quite possible that a focused pursuit of the same approach has the potential for providing a key to solving several other problems arising out of electromechanical systems that are otherwise intractable.

References

1. Narendra, K. S., "Parameter adaptive control—The End … or The Beginning?," *Proceedings of the 33rd Conference on Decision and Control*. Lake Buena Vista, Florida, December 1994.
2. Slotine, J. E. and Li, W., *Applied Nonlinear Control*. Prentice-Hall, Englewood Cliffs, NJ, 1991.
3. Khalil, H. K., *Nonlinear Systems*. Macmillan, New York, NY, 1992.
4. Sastry, S. and Bodson, M., *Adaptive Control: Stability, Convergence and Robustness*. Prentice-Hall, 1989.
5. Tao, G., "A simple alternative proof to the Barbalat Lemma," *IEEE Transactions on Automatic Control*, Vol. 42, No. 5, May 1997, p. 698.
6. Narendra, K. S. and Annaswamy, A. M., *Stable Adaptive Systems*. Prentice-Hall, 1989.
7. Ioannou, P. A. and Sun, J., *Stable and Robust Adaptive Control*. Prentice-Hall, Upper Saddle River, NJ, 1995, pp. 85–134.
8. Astrom, K. J. and Wittenmark, B., *Adaptive Control*. Addison-Wesley, Reading, MA, 1995.
9. Gantmacher. *The Theory of Matrices,* Vol I. Chelsea Publishing Company, NY, 1977, pp. 353–354.
10. Krstić, M., Kanellakopoulos, I., and Kokotović, P. V., "Transient performance improvement with a new class of adaptive controllers," *Systems & Control Letters,* Vol. 21, 1993, pp. 451–461.

11. Krtić, M., Kanellakopoulos, I., and Kokotović, P. V., "Nonlinear design of adaptive controllers for linear systems," *IEEE Transactions on Automatic Control,* Vol. 39, 1994, pp. 738–752.
12. Kokotovic, P. V., "The joy of feedback: nonlinear and adaptive control," *IEEE Control Systems Magazine,* Vol. 12, No. 3, 1992, pp. 7–17.
13. Schaub, H., Akella, M. R., and Junkins, J. L., "Adaptive control of nonlinear attitude motions realizing linear closed loop dynamics," *Journal of Guidance, Control and Dynamics,* Vol. 24, No. 1, Jan.–Feb. 2001.
14. Akella, M. R., Schaub, H., and Junkins, J. L., "Adaptive realization of linear closed loop tracking dynamics in the presence of large system model errors," *Journal of Astronautical Sciences,* Vol. 48, No. 4, 2000.
15. Besançon, G., "Global output feedback tracking control for a class of Lagrangian systems," *Automatica,* Vol. 36, 2000, pp. 1915–1921.
16. Akella, M. R., "Rigid body attitude tracking without angular velocity feedback," *Systems & Control Letters,* Vol. 42, No. 4, 2001.
17. Lizarralde, F. and Wen, J. T., "Attitude control without angular velocity measurement: a passivity approach," *IEEE Transactions on Automatic Control,* Vol. 41, No. 3, 1996, pp. 468–472.
18. Tsiotras, P., "Further passivity results for the attitude control problem," *IEEE Transactions on Automatic Control,* Vol. 43, No. 11, 1998, pp. 1597–1600.

32

Neural Networks and Fuzzy Systems

Bogdan M. Wilamowski
University of Wyoming

32.1 Neural Networks and Fuzzy Systems

New and better electronic devices have inspired researchers to build intelligent machines operating in a fashion similar to the human nervous system. Fascination with this goal started when McCulloch and Pitts (1943) developed their model of an elementary computing neuron and when Hebb (1949) introduced his *learning rules*. A decade later Rosenblatt (1958) introduced the **perceptron** concept. In the early 1960s Widrow and Holf (1960, 1962) developed intelligent systems such as ADALINE and MADALINE. Nillson (1965) in his book *Learning Machines* summarized many developments of that time. The publication of the Mynsky and Paper (1969) book, with some discouraging results, stopped for some time the fascination with artificial neural networks, and achievements in the mathematical foundation of the **backpropagation** algorithm by Werbos (1974) went unnoticed. The current rapid growth in the area of neural networks started with the Hopfield (1982, 1984) recurrent network, Kohonen (1982) unsupervised training algorithms, and a description of the backpropagation algorithm by Rumelhart et al. (1986).

32.2 Neuron Cell

A biological neuron is a complicated structure, which receives trains of pulses on hundreds of *excitatory* and *inhibitory* inputs. Those incoming pulses are summed with different weights (averaged) during the time period of *latent summation*. If the summed value is higher than a threshold, then the neuron itself is generating a pulse, which is sent to neighboring neurons. Because incoming pulses are summed with time, the neuron generates a pulse train with a higher frequency for higher positive excitation. In other words, if the value of the summed weighted inputs is higher, the neuron generates pulses more frequently. At the same time, each neuron is characterized by the nonexcitability for a certain time after the firing pulse. This so-called *refractory period* can be more accurately described as a phenomenon where after excitation the threshold value increases to a very high value and then decreases gradually with a certain time constant. The refractory period sets soft upper limits on the frequency of the output pulse train. In the biological neuron, information is sent in the form of frequency modulated pulse trains.

This description of neuron action leads to a very complex neuron model, which is not practical. McCulloch and Pitts (1943) show that even with a very simple neuron model, it is possible to build logic and memory circuits. Furthermore, these simple neurons with thresholds are usually more powerful than typical logic gates used in computers. The McCulloch–Pitts neuron model assumes that incoming and outgoing signals may have only binary values 0 and 1. If incoming signals summed through positive or negative weights have a value larger than threshold, then the neuron output is set to 1. Otherwise, it is set to 0.

$$T = \begin{cases} 1, & \text{if } net \geq T \\ 0, & \text{if } net < T \end{cases} \tag{32.1}$$

where T is the threshold and *net* value is the weighted sum of all incoming signals:

$$net = \sum_{i=1}^{n} w_i x_i \tag{32.2}$$

Examples of McCulloch–Pitts neurons realizing OR, AND, NOT, and MEMORY operations are shown in Fig. 32.1. Note that the structure of OR and AND gates can be identical. With the same structure, other logic functions can be realized, as Fig. 32.2 shows.

The perceptron model has a similar structure. Its input signals, the weights, and the thresholds could have any positive or negative values. Usually, instead of using variable threshold, one additional constant input with a negative or positive weight can added to each neuron, as Fig. 32.3 shows. In this case, the

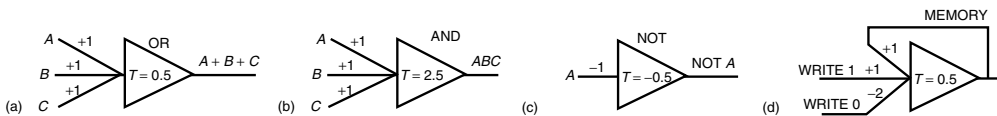



FIGURE 32.1 OR, AND, NOT, and MEMORY operations using networks with McCulloch–Pitts neuron model.

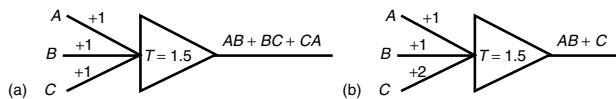



FIGURE 32.2 Other logic function realized with McCulloch–Pitts neuron model.

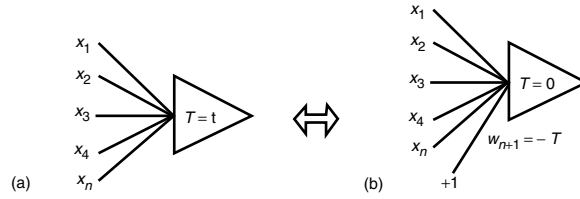



FIGURE 32.3 Threshold implementation with an additional weight and constant input with +1 value: (a) neuron with threshold T, (b) modified neuron with threshold $T = 0$ and additional weight equal to $-T$.

threshold is always set to be zero and the net value is calculated as

$$net = \sum_{i=1}^{n} w_i x_i + w_{n+1} \tag{32.3}$$

where w_{n+1} has the same value as the required threshold and the opposite sign. Single-layer perceptrons were successfully used to solve many pattern classification problems. The hard threshold activation functions are given by

$$o = f(net) = \frac{\mathrm{sgn}(net) + 1}{2} = \begin{cases} 1, & \text{if } net \geq 0 \\ 0, & \text{if } net < 0 \end{cases} \tag{32.4}$$

for **unipolar** neurons and

$$o = f(net) = \mathrm{sgn}(net) = \begin{cases} 1, & \text{if } net \geq 0 \\ -1, & \text{if } net < 0 \end{cases} \tag{32.5}$$

for **bipolar** neurons. For these types of neurons, most of the known training algorithms are able to adjust weights only in single-layer networks.

Multilayer neural networks usually use continuous activation functions, either unipolar

$$o = f(net) = \frac{1}{1 + \exp(-\lambda net)} \tag{32.6}$$

or bipolar

$$o = f(net) = \tanh(0.5\lambda net) = \frac{2}{1 + \exp(-\lambda net)} - 1 \tag{32.7}$$

These continuous activation functions allow for the gradient-based training of multilayer networks. Typical activation functions are shown in Fig. 32.4. In the case when neurons with additional threshold input are used (Fig. 32.3(b)), the λ parameter can be eliminated from Eqs. (32.6) and (32.7) and the steepness of the neuron response can be controlled by the weight scaling only. Therefore, there is no real need to use neurons with variable gains.

Note, that even neuron models with continuous activation functions are far from an actual biological neuron, which operates with frequency modulated pulse trains.

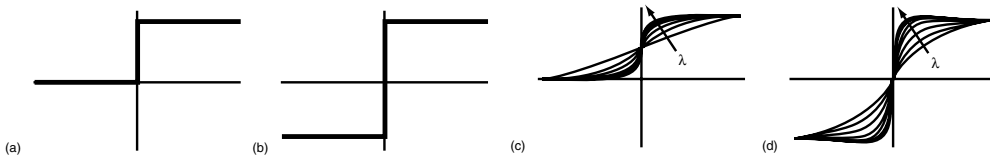



FIGURE 32.4 Typical activation functions: (a) hard threshold unipolar, (b) hard threshold bipolar, (c) continuous unipolar, (d) continuous bipolar.

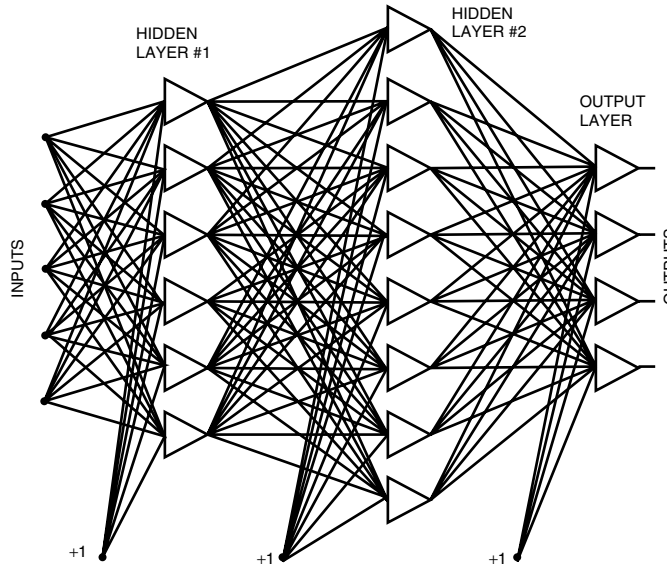



FIGURE 32.5 An example of the three-layer feedforward neural network, which is sometimes known also as the backpropagation network.

32.3 Feedforward Neural Networks

Feedforward neural networks allow only one-directional signal flow. Furthermore, most feedforward neural networks are organized in layers. An example of the three-layer feedforward neural network is shown in Fig. 32.5. This network consists of input nodes, two *hidden layers*, and an output layer.

A single neuron is capable of separating input patterns into two categories, and this separation is linear. For example, for the patterns shown in Fig. 32.6, the separation line is crossing x_1 and x_2 axes at points x_{10} and x_{20}. This separation can be achieved with a neuron having the following weights: $w_1 = 1/x_{10}$, $w_2 = 1/x_{20}$, and $w_3 = -1$. In general for n dimensions, the weights are

$$w_i = \frac{1}{x_{i0}} \quad \text{for } w_{n+1} = -1 \tag{32.8}$$

One neuron can divide only linearly separated patterns. To select just one region in n-dimensional input space, more than $n + 1$ neurons should be used. If more input clusters are to be selected, then the number of neurons in the input (hidden) layer should be properly multiplied. If the number of neurons in the input (hidden) layer is not limited, then all classification problems can be solved using the three-layer network. An example of such a neural network, classifying three clusters in the two-dimensional space, is shown in Fig. 32.7. Neurons in the first hidden layer create the separation lines between input clusters.

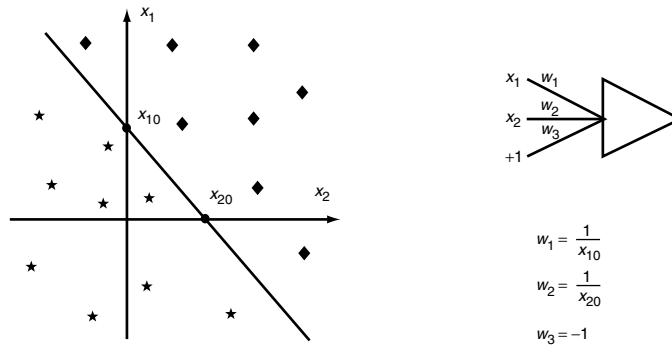



FIGURE 32.6 Illustration of the property of linear separation of patterns in the two-dimensional space by a single neuron.

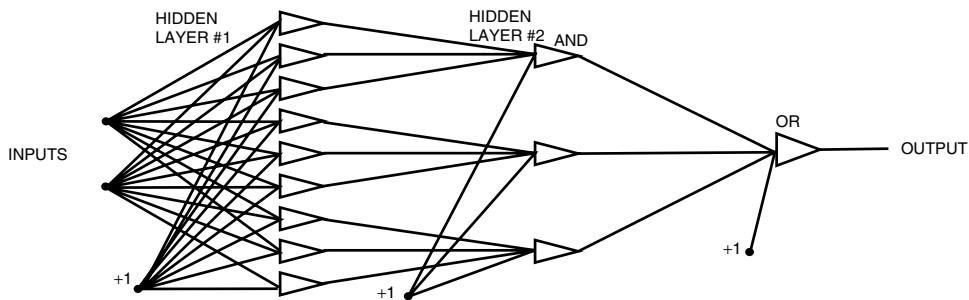



FIGURE 32.7 An example of the three-layer neural network with two inputs for classification of three different clusters into one category. This network can be generalized and can be used for solution of all classification problems.

Neurons in the second hidden layer perform the AND operation, as shown in Fig. 32.1(b). Output neurons perform the OR operation, as shown in Fig. 32.1(a), for each category. The linear separation property of neurons makes some problems especially difficult for neural networks, such as exclusive OR, parity computation for several bits, or to separate patterns laying on two neighboring spirals.

The feedforward neural network is also used for nonlinear transformation (mapping) of a multidimensional input variable into another multidimensional variable in the output. In theory, any input–output mapping should be possible if the neural network has enough neurons in hidden layers. (size of output layer is set by the number of outputs required). In practice, this is not an easy task. Presently, there is no satisfactory method to define how many neurons should be used in hidden layers. Usually, this is found by the trial-and-error method. In general, it is known that if more neurons are used, more complicated shapes can be mapped. On the other hand, networks with large numbers of neurons lose their ability for generalization, and it is more likely that such networks will also try to map noise supplied to the input.

32.4 Learning Algorithms for Neural Networks

Similarly to the biological neurons, the weights in artificial neurons are adjusted during a training procedure. Various learning algorithms were developed, and only a few are suitable for multilayer neuron networks. Some use only local signals in the neurons, others require information from outputs; some require a supervisor who knows what outputs should be for the given patterns, and other unsupervised algorithms need no such information. Common learning rules are described in the following sections.

Hebbian Learning Rule

The Hebb (1949) learning rule is based on the assumption that if two neighbor neurons must be activated and deactivated at the same time, then the weight connecting these neurons should increase. For neurons operating in the opposite phase, the weight between them should decrease. If there is no signal correlation, the weight should remain unchanged. This assumption can be described by the formula

$$\Delta w_{ij} = c x_i o_j \tag{32.9}$$

where

w_{ij} = weight from ith to jth neuron,
c = learning constant,
x_i = signal on the ith input,
o_j = output signal.

The training process starts usually with values of all weights set to zero. This learning rule can be used for both soft and hard threshold neurons. Since desired responses of neurons are not used in the learning procedure, this is the **unsupervised learning** rule. The absolute values of the weights are usually proportional to the learning time, which is undesired.

Correlation Learning Rule

The correlation learning rule is based on a similar principle as the Hebbian learning rule. It assumes that weights between simultaneously responding neurons should be largely positive, and weights between neurons with opposite reaction should be largely negative. Contrary to the Hebbian rule, the correlation rule is the **supervised learning**. Instead of actual response, o_j, the desired response, d_j, is used for the weight change calculation

$$\Delta w_{ij} = c x_i d_j \tag{32.10}$$

This training algorithm usually starts with initialization of weights to zero values.

Instar Learning Rule

If input vectors and weights are normalized, or they have only binary bipolar values (-1 or $+1$), then the *net* value will have the largest positive value when the weights and the input signals are the same. Therefore, weights should be changed only if they are different from the signals

$$\Delta w_i = c(x_i - w_i) \tag{32.11}$$

Note, that the information required for the weight is taken only from the input signals. This is a very local and unsupervised learning algorithm.

Winner Takes All (WTA)

The WTA is a modification of the instar algorithm where weights are modified only for the neuron with the highest *net* value. Weights of remaining neurons are left unchanged. Sometimes this algorithm is modified in such a way that a few neurons with the highest net values are modified at the same time. Although this is an unsupervised algorithm because we do not know what are desired outputs, there is a need for a "judge" or "supervisor" to find a winner with a largest net value. The WTA algorithm, developed by Kohonen (1982), is often used for automatic clustering and for extracting statistical properties of input data.

Outstar Learning Rule

In the outstar learning rule, it is required that weights connected to a certain node should be equal to the desired outputs for the neurons connected through those weights

$$\Delta w_{ij} = c(d_j - w_{ij}) \tag{32.12}$$

where d_j is the desired neuron output and c is the small learning constant, which further decreases during the learning procedure. This is the supervised training procedure because desired outputs must be known. Both instar and outstar learning rules were developed by Grossberg (1969).

Widrow–Hoff LMS Learning Rule

Widrow and Hoff (1960, 1962) developed a supervised training algorithm, which allows training a neuron for the desired response. This rule was derived so the square of the difference between the net and output value is minimized.

$$Error_j = \sum_{p=1}^{P} (net_{jp} - d_{jp})^2 \tag{32.13}$$

where

$Error_j$ = error for jth neuron,
P = number of applied patterns,
d_{jp} = desired output for jth neuron when pth pattern is applied,
net = given by Eq. (32.2).

This rule is also known as the least mean square (LMS) rule. By calculating a derivative of Eq. (32.13) with respect to w_{ij}, a formula for the weight change can be found:

$$\Delta w_{ij} = cx_i \sum_{p=1}^{P} (d_{jp} - net_{jp}) \tag{32.14}$$

Note that weight change Δw_{ij} is a sum of the changes from each of the individual applied patterns. Therefore, it is possible to correct the weight after each individual pattern was applied. This process is known as *incremental updating; cumulative updating* is when weights are changed after all patterns have been applied. Incremental updating usually leads to a solution faster, but it is sensitive to the order in which patterns are applied. If the learning constant c is chosen to be small, then both methods give the same result. The LMS rule works well for all types of activation functions. This rule tries to enforce the *net* value to be equal to desired value. Sometimes this is not what the oberver is looking for. It is usually not important what the *net* value is, but it is important if the *net* value is positive or negative. For example, a very large *net* value with a proper sign will result in correct output and in large error as defined by Eq. (32.13) and this may be the preferred solution.

Linear Regression

The LMS learning rule requires hundreds or thousands of iterations, using formula (32.14), before it converges to the proper solution. Using the linear regression rule, the same result can be obtained in only one step.

Considering one neuron and using vector notation for a set of the input patterns **X** applied through weight vector **w**, the vector of *net* values **net** is calculated using

$$\mathbf{Xw} = \mathbf{net} \tag{32.15}$$

where

$\mathbf{X}$ = rectangular array $(n + 1) \times p$,
n = number of inputs,
p = number of patterns.

Note that the size of the input patterns is always augmented by one, and this additional weight is responsible for the threshold (see Fig. 32.3(b)). This method, similar to the LMS rule, assumes a linear activation function, and so the *net* values **net** should be equal to desired output values **d**

$$\mathbf{Xw} = \mathbf{d} \tag{32.16}$$

Usually $p > n + 1$, and the preceding equation can be solved only in the least mean square error sense. Using the vector arithmetic, the solution is given by

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{d} \tag{32.17}$$

When traditional method is used, the set of p equations with $n + 1$ unknowns, Eq. (32.16), has to be converted to the set of $n + 1$ equations with $n + 1$ unknowns

$$\mathbf{Yw} = \mathbf{z} \tag{32.18}$$

where elements of the **Y** matrix and the **z** vector are given by

$$y_{ij} = \sum_{p=1}^{P} x_{ip}x_{jp}, \qquad z_i = \sum_{p=1}^{P} x_{ip}d_p \tag{32.19}$$

Weights are given by Eq. (32.17) or they can be obtained by a solution of Eq. (32.18).

Delta Learning Rule

The LMS method assumes linear activation function **net** = **o**, and the obtained solution is sometimes far from optimum, as is shown in Fig. 32.8 for a simple two-dimensional case, with four patterns belonging to two categories. In the solution obtained using the LMS algorithm, one pattern is misclassified. If error

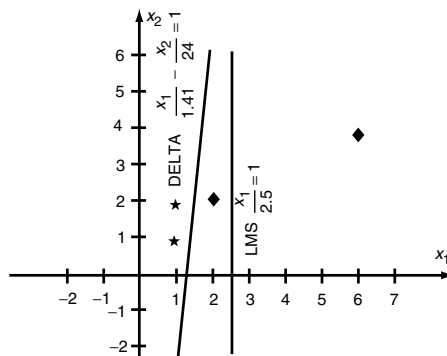



FIGURE 32.8 An example with a comparison of results obtained using LMS and delta training algorithms. Note that LMS is not able to find the proper solution.

is defined as

$$Error_j = \sum_{p=1}^{P} (o_{jp} - d_{jp})^2 \tag{32.20}$$

then the derivative of the error with respect to the weight w_{ij} is

$$\frac{d\ Error_j}{dw_{ij}} = 2\sum_{p=1}^{P} (o_{jp} - d_{jp}) \frac{df\ (net_{jp})}{d\ net_{jp}} x_i \tag{32.21}$$

since $o = f(net)$ and the *net* is given by Eq. (32.2). Note that this derivative is proportional to the derivative of the activation function $f'(net)$. Thus, this type of approach is possible only for continuous activation functions and this method cannot be used with hard activation functions (32.4) and (32.5). In this respect the LMS method is more general. The derivatives' most common continuous activation functions are

$$f' = o(1 - o) \tag{32.22}$$

for the unipolar, Eq. (32.6), and

$$f' = 0.5(1 - o^2) \tag{32.23}$$

for the bipolar, Eq. (32.7).

Using the cumulative approach, the neuron weight w_{ij} should be changed with a direction of gradient

$$\Delta w_{ij} = cx_i \sum_{p=1}^{P} (d_{jp} - o_{jp}) f'_{jp} \tag{32.24}$$

In case of the incremental training for each applied pattern

$$\Delta w_{ij} = cx_i f'_j (d_j - o_j) \tag{32.25}$$

the weight change should be proportional to input signal x_i, to the difference between desired and actual outputs $d_{jp} - o_{jp}$, and to the derivative of the activation function f'_{jp}. Similar to the LMS rule, weights can be updated in both the incremental and the cumulative methods. In comparison to the LMS rule, the delta rule always leads to a solution close to the optimum. As it is illustrated in Fig. 32.8, when the delta rule is used, all four patterns are classified correctly.

Error Backpropagation Learning

The delta learning rule can be generalized for multilayer networks. Using an approach similiar to the delta rule, the gradient of the global error can be computed with respect to each weight in the network. Interestingly,

$$\Delta w_{ij} = cx_i f'_j E_j \tag{32.26}$$

where

c = learning constant,
x_i = signal on the ith neuron input,
f'_j = derivative of activation function.

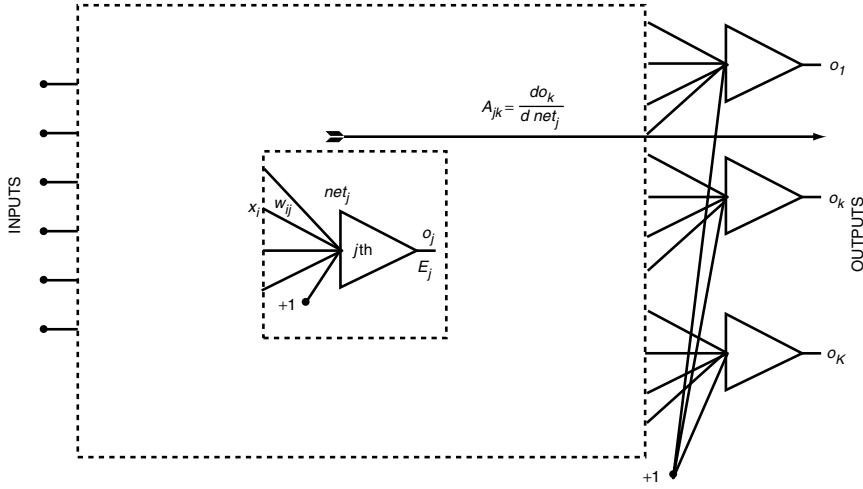



FIGURE 32.9 Illustration of the concept of gain computation in neural networks.

The cumulative error E_j on neuron output is given by

$$E_j = \frac{1}{f_j'} \sum_{k=1}^{K} (o_k - d_k) A_{jk} \tag{32.27}$$

where K is the number of network outputs and A_{jk} is the small signal gain from the input of the jth neuron to the kth network output, as Fig. 32.9 shows. The calculation of the backpropagating error starts at the output layer and cumulative errors are calculated layer by layer to the input layer. This approach is not practical from the point of view of hardware realization. Instead, it is simpler to find signal gains from the input of the jth neuron to each of the network outputs (Fig. 32.9). In this case, weights are corrected using

$$\Delta w_{ij} = c x_i \sum_{k=1}^{K} (o_k - d_k) A_{jk} \tag{32.28}$$

Note that this formula is general, regardless of whether the neurons are arranged in layers or not. One way to find gains A_{jk} is to introduce an incremental change on the input of the jth neuron and observe the change in the kth network output. This procedure requires only forward signal propagation, and it is easy to implement in a hardware realization. Another possible way is to calculate gains through each layer and then find the total gains as products of layer gains. This procedure is equally or less computationally intensive than a calculation of cumulative errors in the error backpropagation algorithm.

The backpropagation algorithm has a tendency for oscillation. To smooth the process, the weights increment Δw_{ij} can be modified according to Rumelhart, Hinton, and Wiliams (1986):

$$w_{ij}(n+1) = w_{ij}(n) + \Delta w_{ij}(n) + \alpha \Delta w_{ij}(n-1) \tag{32.29}$$

or according to Sejnowski and Rosenberg (1987),

$$w_{ij}(n+1) = w_{ij}(n) + (1-\alpha)\Delta w_{ij}(n) + \alpha \Delta w_{ij}(n-1) \tag{32.30}$$

where α is the momentum term.

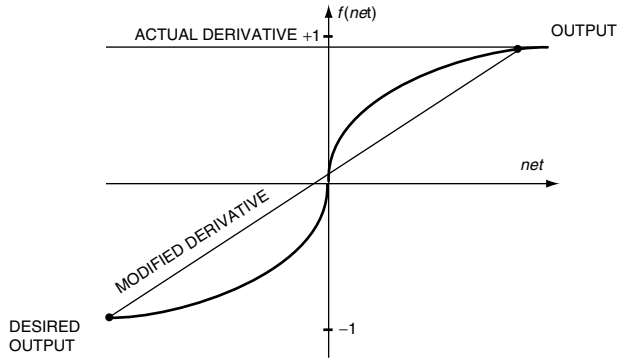



FIGURE 32.10 Illustration of the modified derivative calculation for faster convergency of the error backpropagation algorithm.

The backpropagation algorithm can be significantly sped up, when, after finding components of the gradient, weights are modified along the gradient direction until a minimum is reached. This process can be carried on without the necessity of a computationally intensive gradient calculation at each step. The new gradient components are calculated once a minimum is obtained in the direction of the previous gradient. This process is only possible for cumulative weight adjustment. One method of finding a minimum along the gradient direction is the *tree step process* of finding error for three points along gradient direction and then, using a parabola approximation, jump directly to the minimum. The fast learning algorithm using the described approach was proposed by Fahlman (1988) and is known as the *quickprop*.

The backpropagation algorithm has many disadvantages, which lead to very slow convergency. One of the most painful is that in the backpropagation algorithm, the learning process almost perishes for neurons responding with the maximally wrong answer. For example, if the value on the neuron output is close to +1 and desired output should be close to −1, then the neuron gain $f'(net) \approx 0$ and the error signal cannot backpropagate, and so the learning procedure is not effective. To overcome this difficulty, a modified method for derivative calculation was introduced by Wilamowski and Torvik (1993). The derivative is calculated as the slope of a line connecting the point of the output value with the point of the desired value, as shown in Fig. 32.10.

$$f_{\text{modif}} = \frac{o_{\text{desired}} - o_{\text{actual}}}{net_{\text{desired}} - net_{\text{actual}}} \tag{32.31}$$

Note that for small errors, Eq. (32.31) converges to the derivative of activation function at the point of the output value. With an increase of system dimensionality, the chances for local minima decrease. It is believed that the described phenomenon, rather than a trapping in local minima, is responsible for convergency problems in the error backpropagation algorithm.

32.5 Special Feedforward Networks

The multilayer backpropagation network, as shown in Fig. 32.5, is a commonly used feedforward network. This network consists of neurons with the sigmoid type continuous activation function presented in Figs. 32.4(c) and 32.4(d). In most cases, only the one hidden layer is required, and the number of neurons in the hidden layer are chosen to be proportional to the problem complexity. The number of neurons in the hidden layer is usually found by a trial-and-error process. The training process starts with all weights randomized to small values, and the error backpropagation algorithm is used to find a solution. When the learning process does not converge, the training is repeated with a new set of randomly chosen weights.

Nguyen and Widrow (1990) proposed an experimental approach for the two-layer network weight initialization. In the second layer, weights are randomly chosen in the range from –0.5 to +0.5. In the first layer, initial weights are calculated from

$$w_{ij} = \frac{\beta z_{ij}}{\|z_j\|}, \qquad w_{(n+1)j} = \text{random}(-\beta, +\beta) \tag{32.32}$$

where z_{ij} is the random number from –0.5 to +0.5 and the scaling factor β is given by

$$\beta = 0.7P^{1/N} \tag{32.33}$$

where n is the number of inputs and N is the number of hidden neurons in the first layer. This type of weight initialization usually leads to faster solutions.

For adequate solutions with backpropagation networks, typically many tries are required with different network structures and different initial random weights. It is important that the trained network gains a generalization property. This means that the trained network also should be able to handle correctly patterns that were not used for training. Therefore, in the training procedure, often some data are removed from the training patterns and then these patterns are used for verification. The results with backpropagation networks often depend on luck. This encouraged researchers to develop feedforward networks, which can be more reliable. Some of those networks are described in the following sections.

Functional Link Network

One-layer neural networks are relatively easy to train, but these networks can solve only linearly separated problems. One possible solution for nonlinear problems was presented by Nilsson (1965) and was then elaborated by Pao (1989) using the functional link network shown in Fig. 32.11. Using nonlinear terms with initially determined functions, the actual number of inputs supplied to the one-layer neural network is increased. In the simplest case, nonlinear elements are higher order terms of input patterns. Note that the functional link network can be treated as a one-layer network, where additional input data are generated off-line using nonlinear transformations. The learning procedure for one-layer is easy and fast. Figure 32.12 shows an X OR problem solved using functional link networks. Note that when the functional link approach is used, this difficult problem becomes a trivial one. The problem with the functional link network is that proper selection of nonlinear elements is not an easy task. In many practical cases, however, it is not difficult to predict what kind of transformation of input data may linearize the problem, and so the functional link approach can be used.

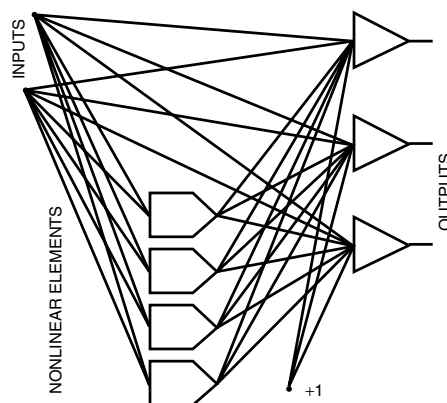



FIGURE 32.11 The functional link network.

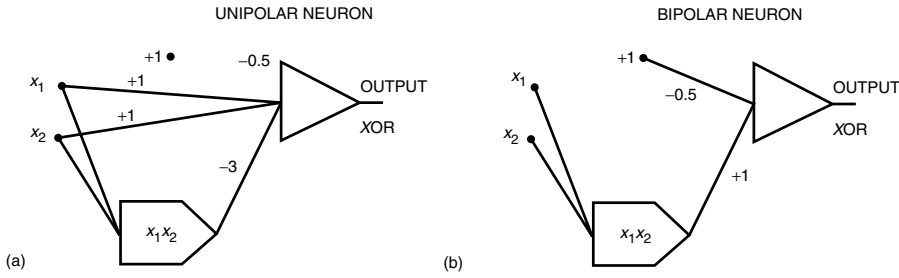



FIGURE 32.12 Functional link networks for solution of the *X* OR problem: (a) using unipolar signals, (b) using bipolar signals.

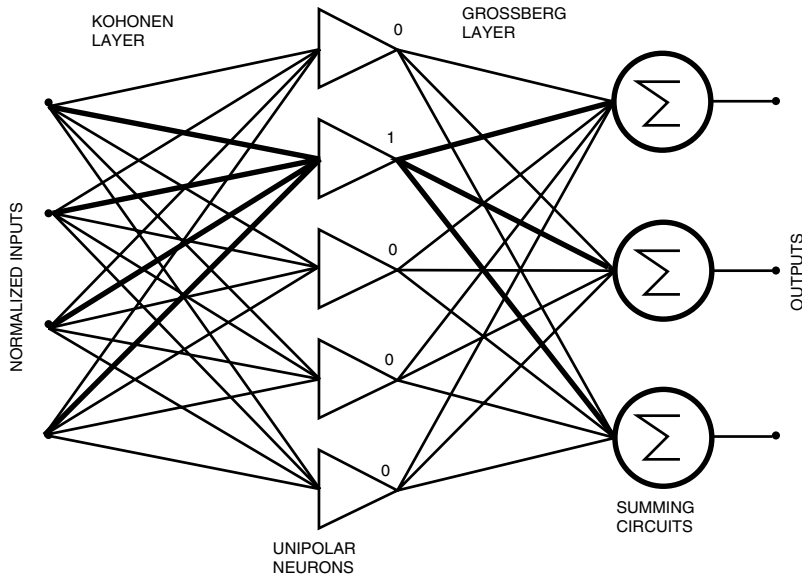



FIGURE 32.13 The counterpropagation network.

Feedforward Version of the Counterpropagation Network

The *counterpropagation network* was originally proposed by Hecht-Nilsen (1987). In this section a modified feedforward version as described by Zurada (1992) is discussed. This network, which is shown in Fig. 32.13, requires numbers of hidden neurons equal to the number of input patterns, or more exactly, to the number of input clusters. The first layer is known as the Kohonen layer with unipolar neurons. In this layer only one neuron, the winner, can be active. The second is the Grossberg outstar layer. The Kohonen layer can be trained in the unsupervised mode, but that need not be the case. When binary input patterns are considered, the input weights must be exactly equal to the input patterns. In this case,

$$net = \mathbf{x}^t\mathbf{w} = [n - 2HD(\mathbf{x}, \mathbf{w})] \tag{32.34}$$

where

n = number of inputs,
$\mathbf{w}$ = weights,
$\mathbf{x}$ = input vector,
$HD(\mathbf{w}, \mathbf{x})$ = *Hamming distance* between input pattern and weights.

For a neuron in the input layer to be reacting just for the stored pattern, the threshold value for this neuron should be

$$w_{(n+1)} = -(n-1) \tag{32.35}$$

If it is required that the neuron must also react for similar patterns, then the threshold should be set to $w_{n+1} = -[n-(1+HD)]$, where HD is the Hamming distance defining the range of similarity. Since for a given input pattern only one neuron in the first layer may have the value of 1 and remaining neurons have 0 values, the weights in the output layer are equal to the required output pattern.

The network, with unipolar activation functions in the first layer, works as a lookup table. When the linear activation function (or no activation function at all) is used in the second layer, then the network also can be considered as an analog memory. For the address applied to the input as a binary vector, the stored set of analog values, as weights in the second layer, can be accurately recovered. The feedforward counterpropagation network may also use analog inputs, but in this case all input data should be normalized,

$$\mathbf{w}_i = \hat{\mathbf{x}}_i = \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|} \tag{32.36}$$

The counterpropagation network is very easy to design. The number of neurons in the hidden layer is equal to the number of patterns (clusters). The weights in the input layer are equal to the input patterns, and the weights in the output layer are equal to the output patterns. This simple network can be used for rapid prototyping. The counterpropagation network usually has more hidden neurons than required. However, such an excessive number of hidden neurons are also used in more sophisticated feedforward networks such as the *probabilistic neural network* (PNN) Specht (1990) or the *general regression neural networks* (GRNN) Specht (1992).

WTA Architecture

The winner take all (WTA) network was proposed by Kohonen (1988). This is basically a one-layer network used in the unsupervised training algorithm to extract a statistical property of the input data, Fig. 32.14(a). At the first step, all input data are normalized so that the length of each input vector is the same and, usually, equal to unity, Eq. (32.36). The activation functions of neurons are unipolar and continuous. The learning process starts with a weight initialization to small random values. During the learning process the weights are changed only for the neuron with the highest value on the output—the winner:

$$\Delta\mathbf{w}_w = c(\mathbf{x} - \mathbf{w}_w) \tag{32.37}$$

where

$\mathbf{w}_w$ = weights of the winning neuron,
$\mathbf{x}$ = input vector,
c = learning constant.

Usually, this single-layer network is arranged into a two-dimensional layer shape, as shown in Fig. 32.14(b). The hexagonal shape is usually chosen to secure strong interaction between neurons. Also, the algorithm is modified in such a way that not only the winning neuron but also neighboring neurons are allowed for the weight change. At the same time, the learning constant c in Eq. (32.37) decreases with the distance from the winning neuron. After such an unsupervised training procedure, the Kohonen layer is able to organize data into clusters. Output of the Kohonen layer is then connected to the one- or two-layer feedforward network with the error backpropagation algorithm. This initial data organization in the WTA layer usually leads to rapid training of the following layer or layers.

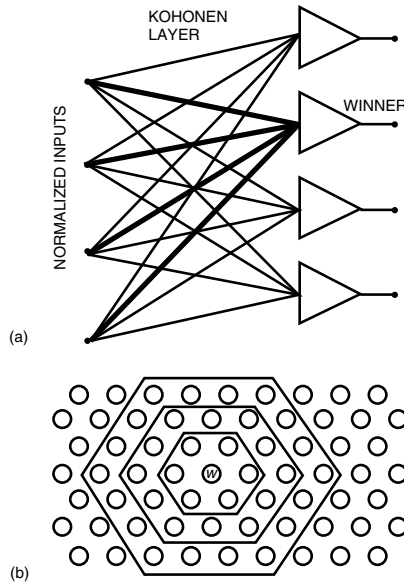



FIGURE 32.14 A winner take all architecture for cluster extracting in the unsupervised training mode: (a) network connections, (b) single-layer network arranged into a hexagonal shape.

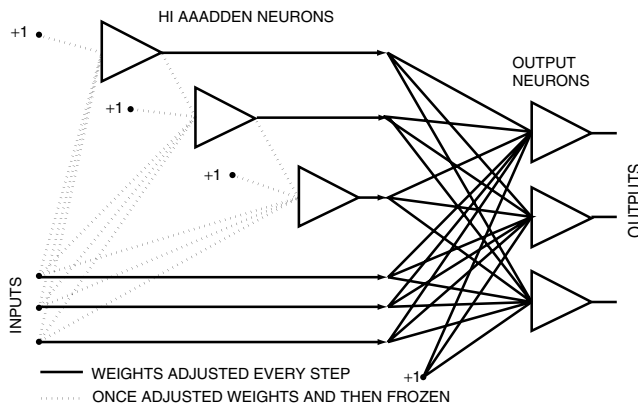



FIGURE 32.15 The cascade correlation architecture.

Cascade Correlation Architecture

The cascade correlation architecture was proposed by Fahlman and Lebiere (1990). The process of network building starts with a one-layer neural network and hidden neurons are added as needed. The network architecture is shown in Fig. 32.15. In each training step, a new hidden neuron is added and its weights are adjusted to maximize the magnitude of the correlation between the new hidden neuron output and the residual error signal on the network output to be eliminated. The correlation parameter S must be maximized:

$$S = \sum_{o=1}^{O} \left| \sum_{p=1}^{P} (V_p - \bar{V})(E_{po} - \bar{E}_o) \right| \tag{32.38}$$

where

O = number of network outputs,
P = number of training patterns,
V_p = output on the new hidden neuron,
E_{po} = error on the network output.

$\bar{V}$ and $\bar{E}_o$ are average values of V_p and E_{po}, respectively. By finding the gradient, $\delta S/\delta w_i$, the weight, adjustment for the new neuron can be found as

$$\Delta w_i = \sum_{o=1}^{O}\sum_{p=1}^{P}\sigma_o(E_{po} - \bar{E}_o)f'_p x_{ip} \tag{32.39}$$

where

σ_o = sign of the correlation between the new neuron output value and network output,
f'_p = derivative of activation function for pattern p,
x_{ip} = input signal.

The output neurons are trained using the delta or quickprop algorithms. Each hidden neuron is trained just once and then its weights are frozen. The network learning and building process is completed when satisfactory results are obtained.

Radial Basis Function Networks

The structure of the radial basis network is shown in Fig. 32.16. This type of network usually has only one hidden layer with special neurons. Each of these neurons responds only to the inputs signals close to the stored pattern. The output signal h_i of the ith hidden neuron is computed using formula

$$h_i = \exp\left(-\frac{\|\mathbf{x} - \mathbf{s}_i\|^2}{2\sigma^2}\right) \tag{32.40}$$

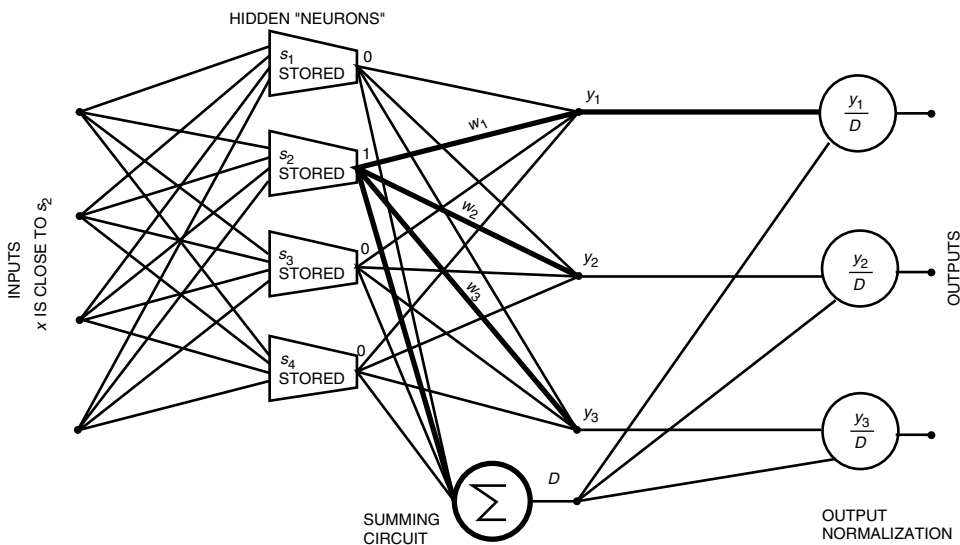



FIGURE 32.16 A typical structure of the radial basis function network.

where

$\mathbf{x}$ = input vector,

$\mathbf{s}_i$ = stored pattern representing the center of the i cluster,

σ_i = radius of the cluster.

Note that the behavior of this "neuron" significantly differs form the biological neuron. In this "neuron," excitation is not a function of the weighted sum of the input signals. Instead, the distance between the input and stored pattern is computed. If this distance is zero, the neuron responds with a maximum output magnitude equal to one. This neuron is capable of recognizing certain patterns and generating output signals that are functions of a similarity. Features of this neuron are much more powerful than a neuron used in the backpropagation networks. As a consequence, a network made of such neurons is also more powerful.

If the input signal is the same as a pattern stored in a neuron, then this neuron responds with 1 and remaining neurons have 0 on the output, as is illustrated in Fig. 32.16. Thus, output signals are exactly equal to the weights coming out from the active neuron. This way, if the number of neurons in the hidden layer is large, then any input–output mapping can be obtained. Unfortunately, it may also happen that for some patterns several neurons in the first layer will respond with a nonzero signal. For a proper approximation, the sum of all signals from the hidden layer should be equal to one. To meet this requirement, output signals are often normalized, as shown in Fig. 32.16.

The radial-based networks can be designed or trained. Training is usually carried out in two steps. In the first step, the hidden layer is usually trained in the unsupervised mode by choosing the best patterns for cluster representation. An approach similar to that used in the WTA architecture can be used. Also in this step, radii σ_i must be found for a proper overlapping of clusters.

The second step of training is the error backpropagation algorithm carried on only for the output layer. Since this is a supervised algorithm for one layer only, the training is very rapid, 100–1000 times faster than in the backpropagation multilayer network. This makes the radial basis-function network very attractive. Also, this network can be easily modeled using computers; however, its hardware implementation would be difficult.

32.6 Recurrent Neural Networks

In contrast to feedforward neural networks, with **recurrent networks** neuron outputs can be connected with their inputs. Thus, signals in the network can continuously circulate. Until recently, only a limited number of recurrent neural networks were described.

Hopfield Network

The single-layer recurrent network was analyzed by Hopfield (1982). This network, shown in Fig. 32.17, has unipolar hard threshold neurons with outputs equal to 0 or 1. Weights are given by a symmetrical square matrix $\mathbf{W}$ with zero elements ($w_{ij} = 0$ for $i = j$) on the main diagonal. The stability of the system is usually analyzed by means of the *energy function*

$$E = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} W_{ij}v_i v_j \tag{32.41}$$

It has been proved that during signal circulation the energy E of the network decreases and the system converges to the stable points. This is especially true when the values of system outputs are updated in the asynchronous mode. This means that at a given cycle, only one random output can be changed to the required values. Hopfield also proved that those stable points, which the system converges, can be

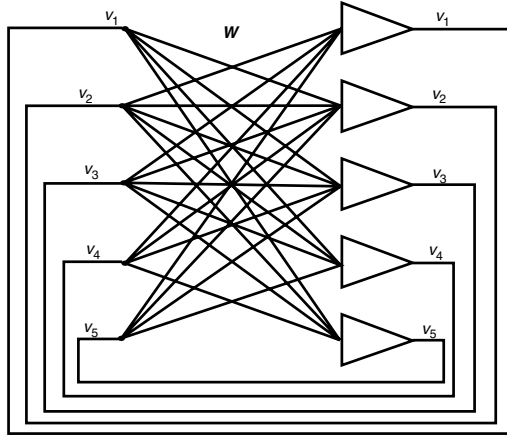



FIGURE 32.17 A Hopfield network or autoassociative memory.

programmed by adjusting the weights using a modified Hebbian rule,

$$\Delta w_{ij} = \Delta w_{ji} = (2v_i - 1)(2v_j - 1)c \tag{32.42}$$

Such memory has limited storage capacity. Based on experiments, Hopfield estimated that the maximum number of stored patterns is $0.15N$, where N is the number of neurons.

Later the concept of energy function was extended by Hopfield (1984) to one-layer recurrent networks having neurons with continuous activation functions. These types of networks were used to solve many optimization and linear programming problems.

Autoassociative Memory

Hopfield (1984) extended the concept of his network to autoassociative memories. In the same network structure as shown in Fig. 32.17, the bipolar hard-threshold neurons were used with outputs equal to −1 or +1. In this network, pattern $\mathbf{s}_m$ are stored into the weight matrix **W** using the autocorrelation algorithm

$$\mathbf{W} = \sum_{m=1}^{M} \mathbf{s}_m \mathbf{s}_m^T - M\mathbf{I} \tag{32.43}$$

where M is the number of stored patterns and **I** is the unity matrix. Note that **W** is the square symmetrical matrix with elements on the main diagonal equal to zero (w_{ji} for $i = j$). Using a modified formula (32.42), new patterns can be added or subtracted from memory. When such memory is exposed to a binary bipolar pattern by enforcing the initial network states, after signal circulation the network will converge to the closest (most similar) stored pattern or to its complement. This stable point will be at the closest minimum of the energy

$$E(\mathbf{v}) = -\frac{1}{2}\mathbf{v}^T \mathbf{W} \mathbf{v} \tag{32.44}$$

Like the Hopfield network, the autoassociative memory has limited storage capacity, which is estimated to be about $M_{\max} = 0.15N$. When the number of stored patterns is large and close to the memory capacity, the network has a tendency to converge to spurious states, which were not stored. These spurious states are additional minima of the energy function.

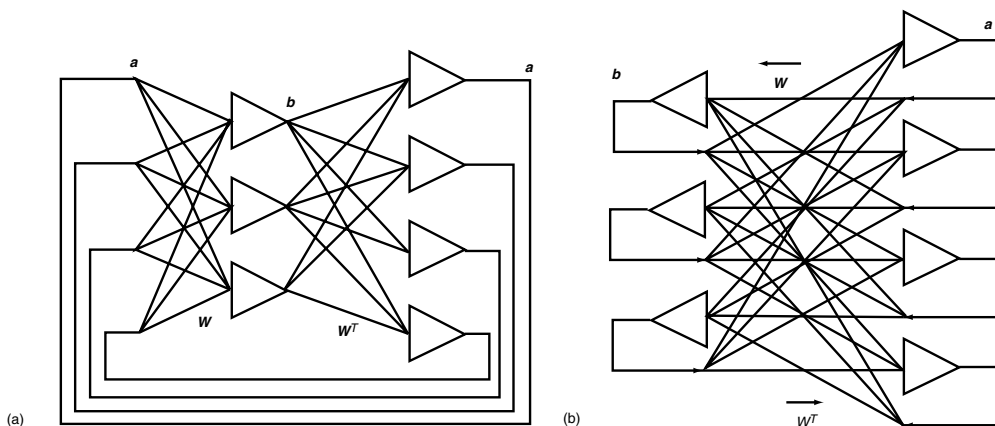



FIGURE 32.18 An example of the bidirectional autoassociative memory: (a) drawn as a two-layer network with circulating signals, (b) drawn as two-layer network with bidirectional signal flow.

Bidirectional Associative Memories (BAMs)

The concept of the autoassociative memory was extended to bidirectional associative memories (BAM) by Kosko (1987, 1988). This memory, shown in Fig. 32.18, is able to associate pairs of the patterns **a** and **b**. This is the two-layer network with the output of the second layer connected directly to the input of the first layer. The weight matrix of the second layer is $\mathbf{W}^T$ and $\mathbf{W}$ for the first layer. The rectangular weight matrix $\mathbf{W}$ is obtained as a sum of the cross-correlational matrices

$$\mathbf{W} = \sum_{m=1}^{M} \mathbf{a}_m \mathbf{b}_m \tag{32.45}$$

where M is the number of stored pairs, and $\mathbf{a}_m$ and $\mathbf{b}_m$ are the stored vector pairs. If the nodes **a** or **b** are initialized with a vector similar to the stored one, then after signal circulations, both stored patterns $\mathbf{a}_m$ and $\mathbf{b}_m$ should be recovered. The BAM has limited memory capacity and memory corruption problems similar to the autoassociative memory. The BAM concept can be extended for association of three or more vectors.

32.7 Fuzzy Systems

The main applications of neural networks are related to the nonlinear mapping of n-dimensional input variables into m-dimensional output variables. Such a function is often required in control systems, where, for specific measured variables, certain control variables must be generated. Another approach for nonlinear mapping of one set of variables into another set of variables is the *fuzzy controller*. The principle of operation of the fuzzy controller significantly differs from neural networks. The block diagram of a fuzzy controller is shown in Fig. 32.19. In the first step, analog inputs are converted into a set of fuzzy variables. In this step, for each analog input, 3–9 fuzzy variables typically are generated. Each fuzzy variable has an analog value between zero and one. In the next step, a fuzzy logic is applied to the input fuzzy variables and a resulting set of output variables is generated. In the last step, known as *defuzzification*, from a set of output fuzzy variables, one or more output analog variables are generated, which are used as control variables.

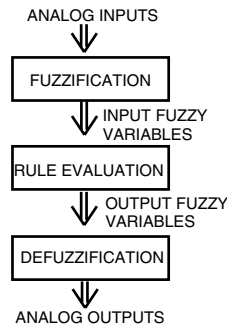



FIGURE 32.19 The block diagram of the fuzzy controller.

FIGURE 32.20 Fuzzification process: (a) typical membership functions for the fuzzification and the defuzzification processes, (b) example of converting a temperature into fuzzy variables.

Fuzzification

The purpose of fuzzification is to convert an analog variable input into a set of fuzzy variables. For higher accuracy, more fuzzy variables will be chosen. To illustrate the fuzzification process, consider that the input variable is the temperature and is coded into five fuzzy variables: cold, cool, normal, warm, and hot. Each fuzzy variable should obtain a value between zero and one, which describes a *degree of association* of the analog input (temperature) within the given fuzzy variable. Sometimes, instead of the term *degree of association*, the term *degree of membership* is used. The process of fuzzification is illustrated in Fig. 32.20. Using Fig. 32.20 we can find the degree of association of each fuzzy variable with the given temperature. For example, for a temperature of 57°F, the following set of fuzzy variables is obtained: [0, 0.5, 0.2, 0, 0], and for $T = 80°F$, it is [0, 0, 0.25, 0.7, 0]. Usually only one or two fuzzy variables have a value other than zero. In the example, trapezoidal functions are used for calculation of the degree of association. Various different functions such as triangular or Gaussian can also be used, as long as the computed value is in the range from zero to one. Each membership function is described by only three or four parameters, which have to be stored in memory.

For proper design of the fuzzification stage, certain practical rules should be used:

- Each point of the input analog variable should belong to at least one and no more than two membership functions.
- For overlapping functions, the sum of two membership functions must not be larger than one. This also means that overlaps must not cross the points of maximum values (ones).
- For higher accuracy, more membership functions should be used. However, very dense functions lead to frequent system reaction and sometimes to system instability.

Rule Evaluation

In contrary to boolean logic where variables can have only binary states, in fuzzy logic all variables may have any values between zero and one. The fuzzy logic consists of the same basic: ∧—AND, ∨—OR, and NOT operators:

$$A \wedge B \wedge C \quad \Rightarrow \quad \min\{A, B, C\}\text{—smallest value of } A \text{ or } B \text{ or } C$$
$$A \vee B \vee C \quad \Rightarrow \quad \max\{A, B, C\}\text{—largest value of } A \text{ or } B \text{ or } C$$
$$\bar{A} \quad \Rightarrow \quad 1\ 1\text{–}A\text{—one minus value of } A$$

(a)

	y_1	y_2	y_3
x_1	z_1	z_1	z_2
x_2	z_1	z_3	z_3
x_3	z_2	z_4	z_4
x_4	z_1	z_2	z_3
x_5	z_1	z_2	z_4

(b)

	y_1	y_2	y_3
x_1	t_{11}	t_{12}	t_{13}
x_2	t_{21}	t_{22}	t_{23}
x_3	t_{31}	t_{32}	t_{33}
x_4	t_{41}	t_{42}	t_{43}
x_5	t_{51}	t_{52}	t_{53}

FIGURE 32.21 Fuzzy tables: (a) table with fuzzy rules, (b) table with the intermediate variables t_{ij}.

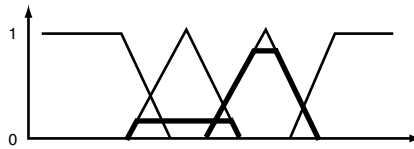



FIGURE 32.22 Illustration of the defuzzification process.

For example, $0.1 \wedge 0.7 \wedge 0.3 = 0.1$, $0.1 \vee 0.7 \vee 0.3 = 0.7$, and $\overline{0.3} = 0.7$. These rules are also known as *Zadeh* AND, OR, and NOT operators (Zadeh, 1965). Note that these rules are true also for classical binary logic.

Fuzzy rules are specified in the *fuzzy table* as it is shown for a given system. Consider a simple system with two analog input variables x and y, and one output variable z. The goal is to design a fuzzy system generating z as $f(x, y)$. After fuzzification, the analog variable x is represented by five fuzzy variables: x_1, x_2, x_3, x_4, x_5 and an analog variable y is represented by three fuzzy variables: y_1, y_2, y_3. Assume that an analog output variable is represented by four fuzzy variables: z_1, z_2, z_3, z_4. The key issue of the design process is to set proper output fuzzy variables z_k for all combinations of input fuzzy variables, as shown in the table in Fig. 32.21. The designer has to specify many rules such as if inputs are represented by fuzzy variables x_i and y_j, then the output should be represented by fuzzy variable z_k. Once the fuzzy table is specified, the fuzzy logic computation proceeds in two steps. First, each field of the fuzzy table is filled with intermediate fuzzy variables t_{ij}, obtained from AND operator $t_{ij} = \min\{x_i, y_j\}$, as shown in Fig. 32.21(b). This step is independent of the required rules for a given system. In the second step, the OR (max) operator is used to compute each output fuzzy variable z_k. In the given example in Fig. 32.21, $z_1 = \max\{t_{11}, t_{12}, t_{21}, t_{41}, t_{51}\}$, $z_2 = \max\{t_{13}, t_{31}, t_{42}, t_{52}\}$, $z_3 = \max\{t_{22}, t_{23}, t_{43},\}$, $z_4 = \max\{t_{32}, t_{34}, t_{53}\}$. Note that the formulas depend on the specifications given in the fuzzy table shown in Fig. 32.21(a).

Defuzzification

As a result of fuzzy rule evaluation, each analog output variable is represented by several fuzzy variables. The purpose of defuzzification is to obtain analog outputs. This can be done by using a membership function similar to that shown in Fig. 32.20. In the first step, fuzzy variables obtained from rule evaluations are used to modify the membership function employing the formula

$$\mu_k^*(z) = \min\{\mu_k(z), z_k\} \tag{32.46}$$

For example, if the output fuzzy variables are 0, 0.2, 0.7, 0.0, then the modified membership functions have shapes shown by the thick line in Fig. 32.22. The analog value of the z variable is found as a *center*

of gravity of modified membership functions $\mu_k^*(z)$,

$$z_{\text{analog}} = \frac{\sum_{k=1}^{n} \int_{-\infty}^{+\infty} \mu_k^*(z) z \, dz}{\sum_{k=1}^{n} \int_{-\infty}^{+\infty} \mu_k^*(z) \, dz} \tag{32.47}$$

In the case where shapes of the output membership functions $\mu_k(z)$ are the same, the equation can be simplified to

$$z_{\text{analog}} = \frac{\sum_{k=1}^{n} z_k zc_k}{\sum_{k=1}^{n} z_k} \tag{32.48}$$

where

n = number of membership function of z_{analog} output variable,
z_k = fuzzy output variables obtained from rule evaluation,
zc_k = analog values corresponding to the center of kth membership function.

Equation (32.47) is usually too complicated to be used in a simple microcontroller based system; therefore, in practical cases, Eq. (32.48) is used more frequently.

Design Example

Consider the design of a simple fuzzy controller for a sprinkler system. The sprinkling time is a function of humidity and temperature. Four membership functions are used for the temperature, three for humidity, and three for the sprinkle time, as shown in Fig. 32.23. Using intuition, the fuzzy table can be developed, as shown in Fig. 32.24(a).

Assume a temperature of 60°F and 70% humidity. Using the membership functions for temperature and humidity, the following fuzzy variables can be obtained for the temperature: [0, 0.2, 0.5, 0], and for the humidity: [0, 0.4, 0.6]. Using the min operator, the fuzzy table can be now filled with temporary fuzzy variables, as shown in Fig. 32.24(b). Note that only four fields have nonzero values. Using fuzzy rules, as shown in Fig. 32.24(a), the max operator can be applied in order to obtain fuzzy output variables: short → $o_1 = \max\{0, 0, 0.2, 0.5, 0\} = 0.5$, medium → $o_2 = \max\{0, 0, 0.2, 0.4, 0\} = 0.4$, long → $o_3 = \max\{0, 0\} = 0$.

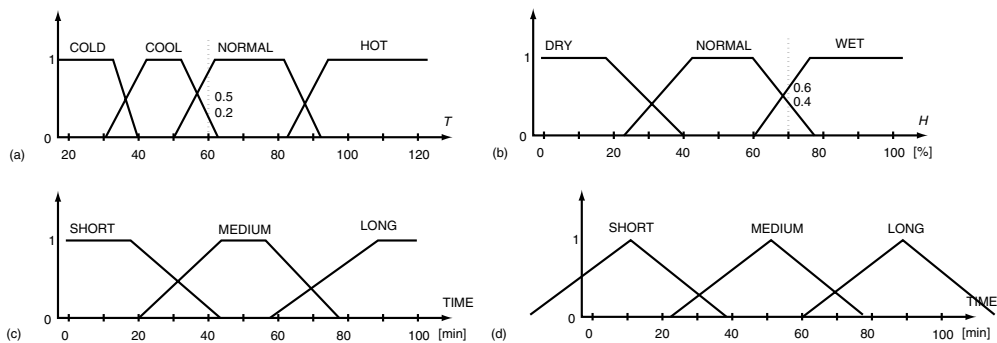



FIGURE 32.23 Membership functions for the presented example: (a) and (b) are membership functions for input variables, (c) and (d) are two possible membership functions for the output variable.

(a)

	DRY	NORMAL	WET
COLD	M	S	S
COOL	M	M	S
WARM	L	M	S
HOT	L	M	S

(b)

	DRY	NORMAL	WET
	0	0.4	0.6
COLD 0	0	0	0
COOL 0.2	0	0.2	0.2
WARM 0.5	0	0.4	0.5
HOT 0	0	0	0

FIGURE 32.24 Fuzzy tables: (a) fuzzy rules for the design example, (b) fuzzy temporary variables for the design example.

Using Eq. (32.47) and Fig. 32.23(c), a sprinkle time of 28 min is determined. When the simplified approach is used with Eq. (32.46) and Fig. 32.23(d), then sprinkle time is 27 min.

32.8 Genetic Algorithms

The success of the artificial neural networks encouraged researchers to search for other patterns in nature to follow. The power of genetics through evolution was able to create such sophisticated machines as the human being. Genetic algorithms follow the evolution process in nature to find better solutions to some complicated problems. The foundations of genetic algorithms are given by Holland (1975) and Goldberg (1989). After initialization, the steps *selection, reproduction with a crossover*, and *mutation* are repeated for each generation. During this procedure, certain strings of symbols, known as chromosomes, evaluate toward a better solution. The genetic algorithm method begins with coding and an initialization. All significant steps of the genetic algorithm will be explained using a simple example of finding a maximum of the function $(\sin^2(x) - 0.5 * x)^2$ with the range of x from 0 to 1.6. Note that in this range, the function has a global maximum at $x = 1.309$, and a local maximum at $x = 0.262$.

Coding and Initialization

At first, the variable x has to be represented as a string of symbols. With longer strings, the process usually converges faster, so the fewer symbols for one string field that are used, the better. Although this string may be the sequence of any symbols, the binary symbols 0 and 1 are usually used. In our example, 6-bit binary numbers are used for coding, having a decimal value of $40x$. The process starts with a random generation of the initial population given in Table 32.1.

Selection and Reproduction

Selection of the best members of the population is an important step in the genetic algorithm. Many different approaches can be used to rank individuals. In this example the ranking function is given. Highest rank has member number 6, and lowest rank has member number 3. Members with higher rank should have higher chances to reproduce. The probability of reproduction for each member can be obtained as a fraction of the sum of all objective function values. This fraction is shown in the last column

TABLE 32.1 Initial Population

String Number	String	Decimal Value	Variable Value	Function Value	Fraction of Total
1	101101	45	1.125	0.0633	0.2465
2	101000	40	1.000	0.0433	0.1686
3	010100	20	0.500	0.0004	0.0016
4	100101	37	0.925	0.0307	0.1197
5	001010	10	0.250	0.0041	0.0158
6	110001	49	1.225	0.0743	0.2895
7	100111	39	0.975	0.0390	0.1521
8	000100	4	0.100	0.0016	0.0062
Total				0.2568	1.0000

of Table 32.1. Note that to use this approach, our objective function should always be positive. If it is not, the proper normalization should be introduced at first.

Reproduction

The numbers in the last column of Table 32.1 show the probabilities of reproduction. Therefore, most likely members numbers 3 and 8 will not be reproduced, and members 1 and 6 may have two or more copies. Using a random reproduction process, the following population, arranged in pairs, could be generated:

$$\begin{array}{llll} 101101 \rightarrow 45 & 110001 \rightarrow 49 & 100101 \rightarrow 37 & 110001 \rightarrow 49 \\ 100111 \rightarrow 39 & 101101 \rightarrow 45 & 110001 \rightarrow 49 & 101000 \rightarrow 40 \end{array}$$

If the size of the population from one generation to another is the same, two parents should generate two children. By combining two strings, two other strings should be generated. The simplest way to do this is to split in half each of the parent strings and exchange substrings between parents. For example, from parent strings 010100 and 100111, the following child strings will be generated: 010111 and 100100. This process is known as the *crossover.* The resultant children are

$$\begin{array}{llll} 101111 \rightarrow 47 & 110101 \rightarrow 53 & 100001 \rightarrow 33 & 110000 \rightarrow 48 \\ 100101 \rightarrow 37 & 101001 \rightarrow 41 & 110101 \rightarrow 53 & 101001 \rightarrow 41 \end{array}$$

In general, the string need not be split in half. It is usually enough if only selected bits are exchanged between parents. It is only important that bit positions are not changed.

Mutation

In the evolutionary process, reproduction is enhanced with mutation. In addition to the properties inherited from parents, offspring acquire some new random properties. This process is known as mutation. In most cases mutation generates low-ranked children, which are eliminated in the reproduction process. Sometimes, however, the mutation may introduce a better individual with a new property. This prevents the process of reproduction from degeneration. In genetic algorithms, mutation usually plays a secondary role. For very high levels of mutation, the process is similar to random pattern generation, and such a searching algorithm is very inefficient. The mutation rate is usually assumed to be at a level well below 1%. In this example, mutation is equivalent to the random bit change of a given pattern. In this simple case, with short strings and a small population, and with a typical mutation rate of 0.1%, the patterns remain practically unchanged by the mutation process. The second generation for this example is shown in Table 32.2.

TABLE 32.2 Population of Second Generation

String Number	String	Decimal Value	Variable Value	Function Value	Fraction of Total
1	010111	47	1.175	0.0696	0.1587
2	100100	37	0.925	0.0307	0.0701
3	110101	53	1.325	0.0774	0.1766
4	010001	41	1.025	0.0475	0.1084
5	100001	33	0.825	0.0161	0.0368
6	110101	53	1.325	0.0774	0.1766
7	110000	48	1.200	0.0722	0.1646
8	101001	41	1.025	0.0475	0.1084
Total				0.4387	1.0000

Note that two identical highest ranking members of the second generation are very close to the solution $x = 1.309$. The randomly chosen parents for the third generation are:

$$\begin{array}{llll} 010111 \rightarrow 47 & 110101 \rightarrow 53 & 110000 \rightarrow 48 & 101001 \rightarrow 41 \\ 110101 \rightarrow 53 & 110000 \rightarrow 48 & 101001 \rightarrow 41 & 110101 \rightarrow 53 \end{array}$$

which produces the following children:

$$\begin{array}{llll} 010101 \rightarrow 21 & 110000 \rightarrow 48 & 110001 \rightarrow 49 & 101101 \rightarrow 45 \\ 110111 \rightarrow 55 & 110101 \rightarrow 53 & 101000 \rightarrow 40 & 110001 \rightarrow 49 \end{array}$$

The best result in the third population is the same as in the second one. By careful inspection of all strings from the second or third generation, it may be concluded that using crossover, where strings are always split in half, the best solution 110100 → 52 will never be reached, regardless of how many generations are created. This is because none of the population in the second generation has a substring ending with 100. For such crossover, a better result can be only obtained due to the mutation process, which may require many generations. Better results in the future generation also can be obtained when strings are split in random places. Another possible solution is that only randomly chosen bits are exchanged between parents.

The genetic algorithm is very rapid, and it leads to a good solution within a few generations. This solution is usually close to global maximum, but not the best.

Defining Terms

Backpropagation: Training technique for multilayer neural networks.
Bipolar neuron: Neuron with output signal between –1 and +1.
Feedforward network: Network without feedback.
Perceptron: Network with hard threshold neurons.
Recurrent network: Network with feedback.
Supervised learning: Learning procedure when desired outputs are known.
Unipolar neuron: Neuron with output signal between 0 and +1.
Unsupervised learning: Learning procedure when desired outputs are unknown.

References

Fahlman, S.E. 1988. Faster-learning variations on backpropagation: An empirical study. *Proceedings of the Connectionist Models Summer School,* D. Touretzky, G. Hinton, and T. Sejnowski, Eds., Morgan Kaufmann, San Mateo, CA.

Fahlman, S.E. and Lebiere, C. 1990. The cascade correlation learning architecture. *Adv. Ner. Inf. Proc. Syst.,* 2, D.S. Touretzky, ed., pp. 524–532. Morgan Kaufmann, Los Altos, CA.

Goldberg, D.E. 1989. *Genetic Algorithm in Search, Optimization and Machine Learning.* Addison–Wesley, Reading, MA.

Grossberg, S. 1969. Embedding fields: a theory of learning with physiological implications. *Journal of Mathematical Psychology* 6:209–239.

Hebb, D.O. 1949. *The Organization of Behavior, a Neuropsychological Theory.* John Wiley, New York.

Hecht-Nielsen, R. 1987. Counterpropagation networks. *Appl. Opt.* 26(23):4979–4984.

Hecht-Nielsen, R. 1988. Applications of counterpropagation networks. *Neural Networks* 1:131–139.

Holland, J.H. 1975. *Adaptation in Natural and Artificial Systems*. Univ. of Michigan Press, Ann Arbor, MI.

Hopfield, J.J. 1982. Neural networks and physical systems with emergent collective computation abilities. *Proceedings of the National Academy of Science* 79:2554–2558.

Hopfield, J.J. 1984. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Science* 81:3088–3092.

Kohonen, T. 1988. The neural phonetic typerater. *IEEE Computer* 27(3):11–22.

Kohonen, T. 1990. The self-organized map. *Proc. IEEE* 78(9):1464–1480.

Kosko, B. 1987. Adaptive bidirectional associative memories. *App. Opt.* 26:4947–4959.

Kosko, B. 1988. Bidirectional associative memories. *IEEE Trans. Sys. Man, Cyb.* 18:49–60.

McCulloch, W.S. and Pitts., W.H., 1943. A logical calculus of the ideas imminent in nervous activity. *Bull. Math. Biophys.* 5:115–133.

Minsky, M. and Papert, S. 1969. *Perceptrons*. MIT Press, Cambridge, MA.

Nilsson, N.J. 1965. *Learning Machines: Foundations of Trainable Pattern Classifiers*. McGraw-Hill, New York.

Nguyen, D. and Widrow, B. 1990. Improving the learning speed of 2-layer neural networks, by choosing initial values of the adaptive weights. *Proceedings of the International Joint Conference on Neural Networks* (San Diego), CA, June.

Pao, Y.H. 1989. *Adaptive Pattern Recognition and Neural Networks*. Addison–Wesley, Reading, MA.

Rosenblatt, F. 1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psych. Rev.* 65:386–408.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. 1986. Learning internal representation by error propagation. *Parallel Distributed Processing*. Vol. 1, pp. 318–362. MIT Press, Cambridge, MA.

Sejnowski, T.J. and Rosenberg, C.R. 1987. Parallel networks that learn to pronounce English text. *Complex Systems* 1:145–168.

Specht, D.F. 1990. Probalistic neural networks. *Neural Networks* 3:109–118.

Specht, D.F. 1992. General regression neural network. *IEEE Trans. Neural Networks* 2:568–576.

Wasserman, P.D. 1989. *Neural Computing Theory and Practice*. Van Nostrand Reinhold, New York.

Werbos, P., 1974. Beyond regression: new tools for prediction and analysis in behavioral sciences. Ph.D. diss., Harvard University.

Widrow, B. and Hoff, M.E. 1960. Adaptive switching circuits. 1960 IRE Western Electric Show and Convention Record, Part 4 (Aug. 23):96–104.

Widrow, B. 1962. Generalization and information storage in networks of adaline Neurons. In *Self-organizing Systems*, M.C. Jovitz, G.T. Jacobi, and G. Goldstein, eds., pp. 435–461. Sparten Books, Washington, D.C.

Wilamowski, M. and Torvik, L., 1993. Modification of gradient computation in the back-propagation algorithm. *ANNIE'93 - Artificial Neural Networks in Engineering*. November 14–17, 1993, St. Louis, Missou.; also in C.H Dagli, ed. 1993. *Intelligent Engineering Systems Through Artificial Neural Networks* Vol. 3, pp. 175–180. ASME Press, New York.

Zadeh, L.A. 1965. Fuzzy sets. *Information and Control* 8:338–353.

Zurada, J. 1992. *Introduction to Artificial Neural Systems*. West Publ.

33

Advanced Control of an Electrohydraulic Axis

Florin Ionescu
University of Applied Sciences

Crina Vlad
Politeknica University of Bucharest

Dragos Arotaritei
Aalborg University Esbjerg

33.1 Introduction

Due to the development of technology in the last few years, robots are seen as advanced mechatronic systems which require knowledge from mechanics, actuators, and control in order to perform very complex tasks. Different kinds of servo-systems, especially electrohydraulic, could be met at the executive level of the robots. Taking into account the most advanced control approaches, this paper deals with the implementation of advanced controllers besides conventional ones which are used in an electrohydraulic system. The considered electrohydraulic system is one of the axes of a robot. These robots possess three or more electrohydraulic axes, which are identical with the axis studied in this chapter.

An electrohydraulic axis whose mathematical model (MM) is described in this chapter presents a multitude of nonlinearities. Conventional controllers are becoming increasingly inappropriate to control the systems with an imprecise model where many nonlinearities are manifested. Therefore, advanced techniques such as neural networks and fuzzy algorithms are deeply involved in the control of such systems. Neural networks, initially proposed by McCulloch and Pitts, Rosenblatt, Widrow, had several

limitations that restricted the domain of applications. An important change took place in the 1980s when Hopfield's theory regarding recurrent neural networks, the model of self-organization developed by Kohonen, and cellular neural networks (Chua) relaunched this research field. The development of some efficient algorithms dedicated specifically to the architecture of neural networks, and the application of these networks in control, represents an interesting area of research in the contemporary world of science.

Fuzzy systems, in conjunction with neural networks, hold an important place in advanced techniques of control. These systems have origins in fuzzy set theory initiated by L. Zadeh. One essential feature of fuzzy systems is the approximate reasoning in which the variables are described in a qualitative manner. Due to the capability of fuzzy systems to deal with imprecise information, they are strongly recommended in order to express knowledge in the form of linguistic rules. In this way, the human operator's knowledge, which is linguistic or numerical, is used to generate the set of fuzzy if-then rules as a basis for a fuzzy controller. A main drawback of fuzzy systems is the difficulty to design them on the basis of a systematic methodology. To overcome this drawback, the learning procedures from neural networks are applied successfully in order to tune the parameters of membership functions.

The merging of these two fields has led to the emergence of neuro-fuzzy systems, which have been applied with promising results in the field of control-engineering. In order to improve dynamic and static performances of the systems characterized by nonlinearities and uncertainties, neuro-fuzzy controllers are used.

The present contribution is organized as follows. An introduction of the electrohydraulic systems with an emphasis on the control of such devices is realized in Section 33.2. Section 33.3 is devoted to the MM of electrohydraulic axes, and the subsequent sections treat the control of electrohydraulic axes through conventional methods (Section 33.4), fuzzy systems (Section 33.5), neural networks (Section 33.6), and neuro-fuzzy techniques (Section 33.7). Conclusions are given in Section 33.8.

33.2 Generalities Concerning ROBI_3, a Cartesian Robot with Three Electrohydraulic Axes

The automated installation, which uses electrohydraulic axes, whose mathematical model is described in section 33.3, is a Cartesian robot named ROBI_3. ROBI_3 has three identical electrohydraulic axes and is built from aluminium profiles [21]. The slides are actuated by hydraulic servoactuators (Rexroth). Slades move on the linear guideways with balls and two recirculation paths. The hydraulic supply installation is placed under the robot table and has a cooling and controlling installation with air. The mechanical structure of the robot is depicted in Fig. 33.1. The three axes of ROBI_3 are identically controlled by the controlling software named TORCH, which runs in Windows. The 32-bit dSPACE controlling hardware endowed with 10 A/D and D/A interfaces is plugged into the PC and serves as the interface between the PC and each of the axes.

An electrohydraulic axis consists of a servovalve and a hydraulic cylinder and has a nonlinear structure. The control system of one axis consists of:

1. the controller represented by a personal computer endowed with a process card;
2. the electrohydraulic converter;
3. the actuator (a linear hydraulic servomotor (LHM));
4. the mechanical process to be controlled, characterized by the slade position;
5. the position transducer.

The control system of robot ROBI_3 is illustrated in Fig. 33.2, where the presence of three electrohydraulic axes, as well as the structure of one axes, identical to the others, can be observed.

The corresponding mathematical model for one axis, on the basis of which the control of the robot is achieved, is described in section 33.3. Through numerical simulations of the three axes of the robot, the necessary mechanical structure interface data is obtained.

Preliminary experiments with driven and controlled variable: position and velocity are made in order to achieve experience (Figs. 33.2 and 33.3). The diagram of a closed-loop position control with

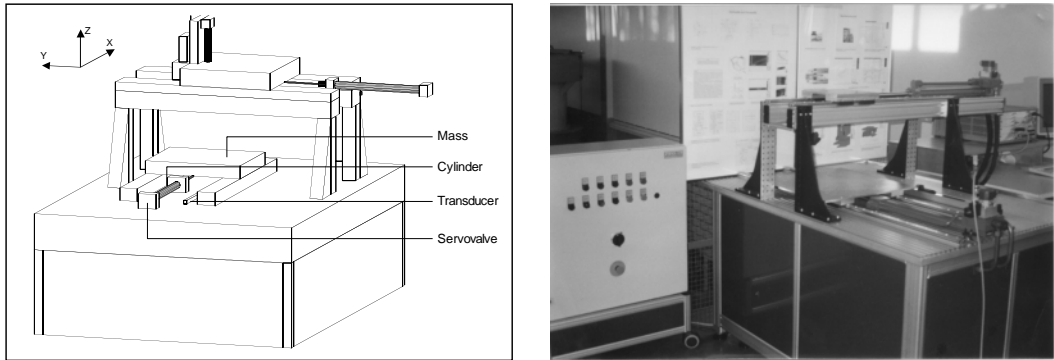



FIGURE 33.1 ROBI_3, a Cartesian robot with three axes [21]. (a) Design; (b) practical allignment.

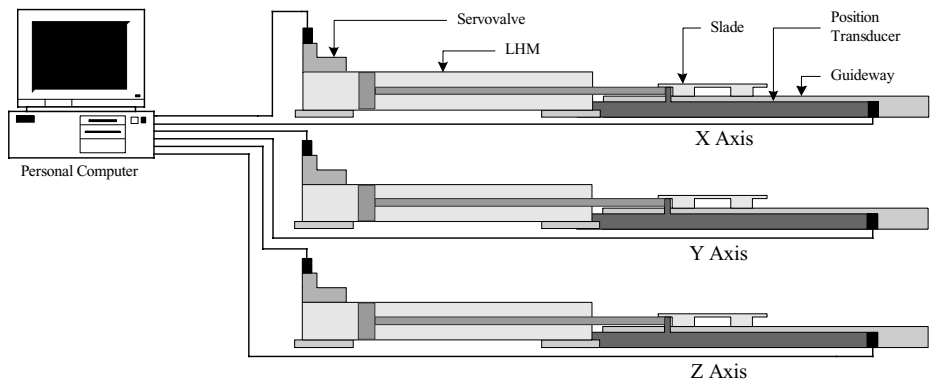



FIGURE 33.2 The control system of the robot [21,24].

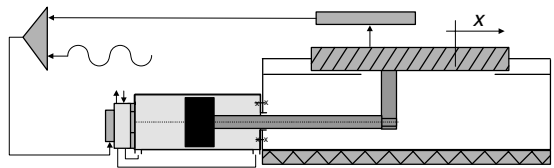



FIGURE 33.3 Diagram of a closed-loop position control with direct measurement.

direct measurement (driven by means of a servovalve) is shown in Fig. 33.3, and the closed-loop position control with indirect measurement at spindle (actuated by means of a servovalve) is shown in Fig. 33.4.

A volumetric Q-rate regulation with constant pressure ($Q \neq$ const; p @ const) is shown in Fig. 33.5. However, this classic model, useful for application, was used only for preliminary results in simulation. One of the main reasons to use this is because we need a well-known mathematical model with well-studied behavior in order to test the controllers (neural and neuro-fuzzy, namely).

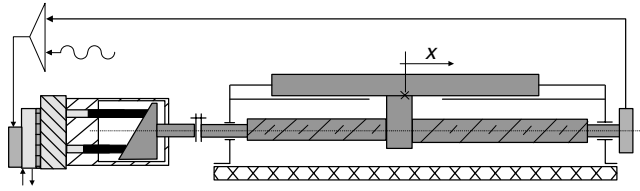



FIGURE 33.4 Closed loop-position control with indirect measurement.

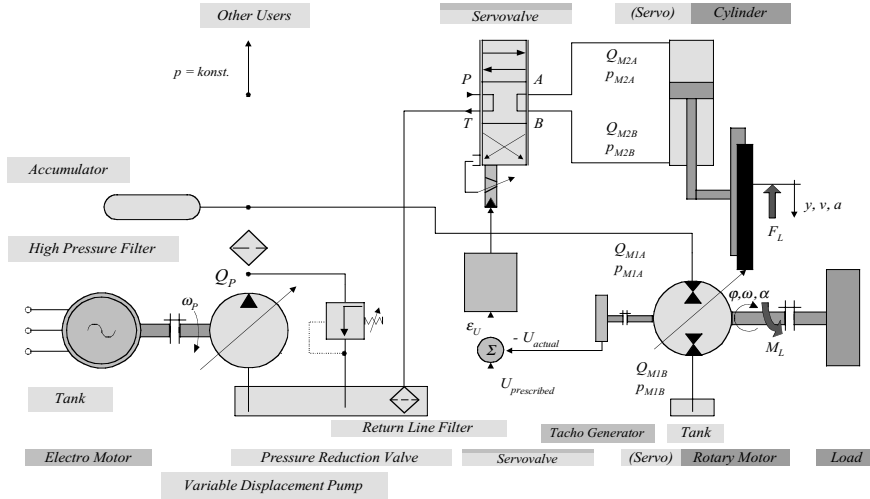



FIGURE 33.5 Volumetric Q-rate regulation with constant pressure.

33.3 Mathematical Model and Simulation of Electrohydraulic Axes

Section 33.3 deals with the analytical findings of the mathematical model (MM) of an electrohydraulic axis, a component part of robot ROBI_3. This method of analysis is advantageous because it offers the possibility to use this MM for other electrohydraulic axes as well, regardless of the different number of stages, and also allows the testing of dynamic performances of the axis at the design level.

In this section, the following were realized:

1. the static models of electrohydraulic system components (servovalve, hydraulic linear motor);
2. the parameters involved in MM, based on constructive and flowing regime characteristics;
3. the nonlinear MM of the proposed electrohydraulic system;
4. the structural scheme of the hydraulic axis in order to simulate its behavior (SIMULINK);
5. the investigations regarding MM, which certify the stability of the system and the fact that the modelled process is a rapid one.

Values of parameters that describe the MM are set based on hydraulic characteristics and on constructive parameters of the considered system.

The Extended Mathematical Model

The studied system consists of a servovalve and an asymmetric motor. In most cases, the control of an electrohydraulic axis is directed towards position control, velocity control, pressure control, or force control. The position control of the axis is studied. The position control loop for the proposed installation is illustrated in Fig. 33.6, together with the denomination of some data.

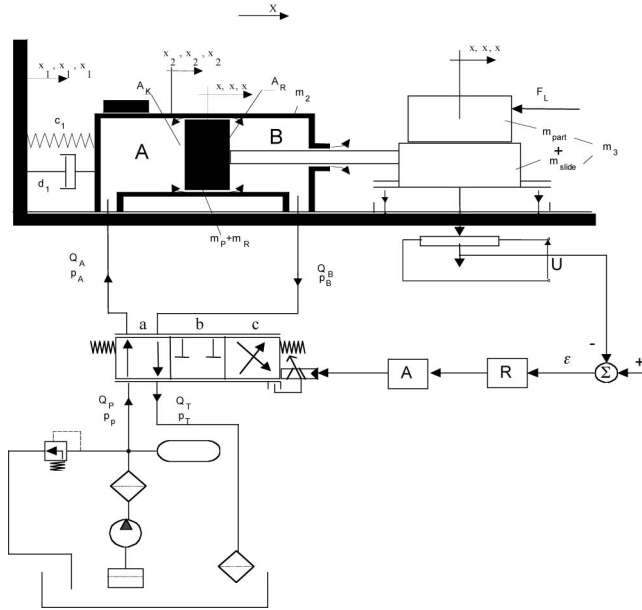



FIGURE 33.6 Control scheme of the servosystem.

Nonlinear Mathematical Model of the Servovalve

As previously mentioned, the servovalve used at the proposed electrohydraulic axis has four paths and three positions. The servovalve has four active control edges at the second hydraulic stage. Three stages could be distinguished by this servovalve: the electromechanical, the mechanohydraulical, and the hydromechanical one. Regarded as a system, a servovalve is complex, with various types of nonlinearities being manifested. Different static and dynamic nonlinearities such as dead zone, jump in origin, saturation, Coulombian and Newtonian frictions with hysteresis, and asymmetry appear in each of these three levels and also in the actuator of the electrohydraulic motor. These were taken into account in the modeling of the servovalve and of the cylinder behavior.

For the studied servovalve, the circulation of the fluid is considered directed from the pump to the admission chamber A (Q_A) and from the discharge chamber B to the reservoir (Q_B). Figure 33.6 presents the control loop where the transducer is placed on the feedback path and the controller R and the amplifier A are on the direct path. Electrical signal (±10 V or ±300 mA) is converted into the displacement x_v of the valve, and in this way in a flow Q, which is transmitted to the linear hydraulic motor.

From the point of view of control characteristics, the number of active edges serves as a method of classification of slide valves [3]. A servovalve with four active control edges was considered. The lightly simplified mathematical model, which describes the functionality of the servovalve, consists of the following equations:

$$u(t) = L \cdot \frac{d\,(t)}{dt} + R \cdot i(t) \tag{33.1}$$

where L [H], the inductance of electrical level; R [W], the resistance of electrical level; $u(t)$ [V], the control voltage; $i(t)$ [A], the control intensity.

$$m \cdot x_v(t) + d \cdot x_v(t) + c \cdot x_v(t) = \sum F \tag{33.2}$$

where m [kg], the mass of valve; d [N/(m/s)], the linearized gradient of viscous friction for the piston of the valve; c[N/m], the coefficient of hydraulic elasticity; x_v [m], the spool displacement; $\sum F$ [N], the

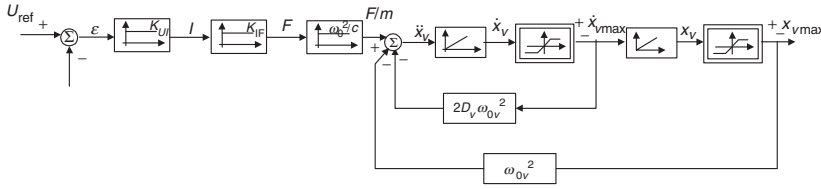



FIGURE 33.7 Nonlinear MM of the first stages of the servovalve.

resulting force, which actuates on the valve spool.

$$\ddot{x}_v(t) + 2 \cdot D_v \cdot \omega_{0v} \cdot \dot{x}_v(t) + \omega_{0v}^2 \cdot x_v(t) = \frac{\Sigma F}{m}$$

$$\ddot{x}_v(t) = k^* - 2 \cdot D_v \cdot \omega_{0v} \cdot \dot{x}_v(t) - \omega_{0v}^2 \cdot x_v(t) \quad \text{where } k^* = \frac{\Sigma F}{m} \tag{33.3}$$

The displacement x_v, obtained based upon the mentioned equations, is implemented in SIMULINK using the scheme from Fig. 33.7. This module of nonlinear MM includes two stages of electrohydraulic axis: the electrohydraulic and the hydromechanical one.

The corresponding equations for the four flows that go through the servovalve are

$$\begin{aligned}
Q_{PA} &= \alpha_D \cdot \pi \cdot D_v \cdot \sqrt{\frac{2}{\rho}} \cdot (x_0 + x_v(t)) \cdot \sqrt{p_P - p_A(t)}, \quad x_v \in [-x_0, x_{\max}] \\
Q_{AT} &= \alpha_D \cdot \pi \cdot D_v \cdot \sqrt{\frac{2}{\rho}} \cdot (x_0 - x_v(t)) \cdot \sqrt{p_A(t) - p_T}, \quad x_v \in [-x_{\max}, x_0] \\
Q_{PB} &= \alpha_D \cdot \pi \cdot D_v \cdot \sqrt{\frac{2}{\rho}} \cdot (x_0 - x_v(t)) \cdot \sqrt{p_P - p_B(t)}, \quad x_v \in [-x_{\max}, x_0] \\
Q_{BT} &= \alpha_D \cdot \pi \cdot D_v \cdot \sqrt{\frac{2}{\rho}} \cdot (x_0 + x_v(t)) \cdot \sqrt{p_B(t) - p_T}, \quad x_v \in [-x_0, x_{\max}]
\end{aligned} \tag{33.4}$$

where Q_{PA} [m^3/s], the flow to the hydraulic motor, from pump to the chamber A of the motor; Q_{AT} [m^3/s], the flow from the chamber A to reservoir; Q_{PB} [m^3/s], the flow from pump to the chamber B of the motor; Q_{BT} [m^3/s], the flow from the chamber B to reservoir; α_D [-], the discharge coefficient; D_v [m], the spool's diameter; x_0 [m], the dimension of the lap of the spool; x_v [m], the spool's displacement; p_A [N/m^2], the fluid pressure in chamber A; p_B [N/m^2], the fluid pressure in chamber B [kg/m^3] fluids density. The flows, which are transmitted to LHM and evacuated from LHM, are Q_A and Q_B, which are computed as following:

$$Q_A = Q_{PA} - Q_{AT}, \qquad Q_B = Q_{BT} - Q_{PB} \tag{33.5}$$

The lap of the spool is considered to be zero ($x_0 = 0$), and, therefore, the static characteristic is linear around the origin and also in the rest. With $Q_0 = \alpha_D \cdot \pi \cdot D_v \cdot \sqrt{2/\rho}$, the flow equations become

$$\begin{aligned}
Q_{PA} &= Q_0 \cdot x_v \cdot \sqrt{p_P - p_A}, \qquad & Q_{AT} &= Q_0 \cdot (-x_v) \cdot \sqrt{p_A - p_T} \\
Q_{PB} &= Q_0 \cdot (-x_v) \cdot \sqrt{p_P - p_B}, \qquad & Q_{BT} &= Q_0 \cdot x_v \cdot \sqrt{p_B - p_T}
\end{aligned} \tag{33.6}$$

Nonlinear Mathematical Model of Linear Hydraulic Motor

The differential equations, based on which the MM of the linear hydraulic motor (LHM) was achieved, are

i. the equation of the dynamic equilibrium of the forces reduced to the motor's rod, and
ii. the equation of movement and flow continuity.

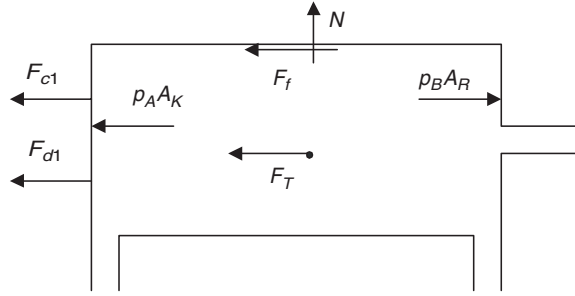



FIGURE 33.8 D'Alembert's principle applied to the cylinder.

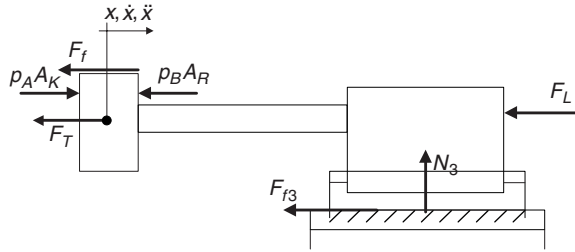



FIGURE 33.9 D'Alembert's principle applied to the rod, piston, and load.

Modeling the LHM, several simplifications (most of them concerning the Coulombian and Newtonian friction) were used. The forces that actuate on the LHM cylinder are depicted in Fig. 33.8:

Applying D'Alembert's principle, the equation of dynamic equilibrium of forces for the cylinder of LHM is

$$\dot{x}_2 = \frac{p_B \cdot A_R - p_A \cdot A_k + c_1 \cdot (x_1 - x_2) + d_1 \cdot (\dot{x}_1 - \dot{x}_2) + c_{\text{fu}} \cdot N \cdot \text{sgn}(\dot{x} - \dot{x}_2) + d_z \cdot (\dot{x} - \dot{x}_2)}{m_2 \cdot s} \qquad (33.7)$$

where c_1 [N/m], the elasticity; d_1 [N/(m/s)], the linearized coefficient of the viscous Newtonian friction in the connection actuators' cylinder and wall; c_{fu} [-], the coefficient of the dry Coulombian friction in the cylinder and rod seals; d_z [N/(m/s)], the coefficient of Newtonian friction in the piston and rod seals; m_2 [kg], the cylinder mass; p_A [N/m^2], the fluid pressure in the admission chamber A of the actuator; p_B [N/m^2], the fluid pressure in the discharge chamber B of the actuator; A_K [m^2], the piston active area in chamber A; A_R [m^2], the piston active area in chamber B; N [N], the normal force, which determines the friction force between piston and cylinder; x [m], the piston displacement; x_1 [m], the wall displacement; and x_2 [m], the cylinder displacement.

The forces acting on the rod, piston, and working element are illustrated in Fig. 33.9

The velocity corresponding to the rod, piston, and mass m_3 (slade, guideway, and loading are considered stiff fastened) is inferred from the equilibrium equation:

$$\dot{x} = \frac{p_A \cdot A_K - p_B \cdot A_R - F_L - c_{\text{fu}} \cdot N \cdot \text{sgn}(\dot{x} - \dot{x}_2) - d_z \cdot (\dot{x} - \dot{x}_2) - c_{\text{fu}3} \cdot N_3 \cdot \text{sgn}(\dot{x} - \dot{x}_1) - d_3 \cdot (\dot{x} - \dot{x}_1)}{(m_p + m_T + m_3) \cdot s} \qquad (33.8)$$

where m_p [kg], the piston mass; m_T [kg], the rod mass; m_3 [kg], load mass (reduced at the rod of piston); $c_{\text{fu}3}$ [-], the coefficient of Coulombian friction between guide ways and slade; d_3 [N/(m/s)], the coefficient of Newtonian friction between guideways and slade; N_3 [N], the normal force which appears between loading and table; F_L [N], the loading force.

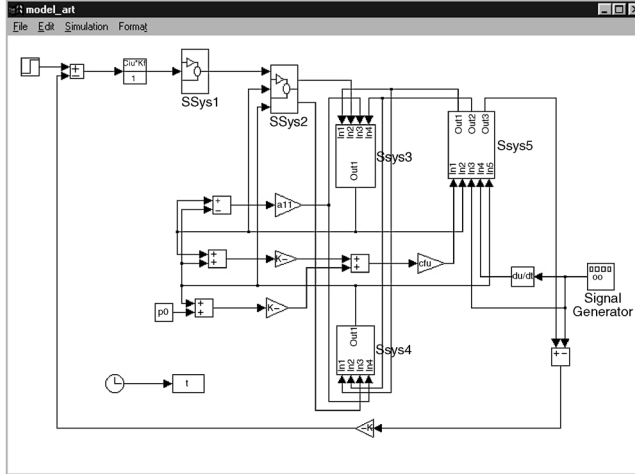



FIGURE 33.10 Electrohydraulic axis implemented in SIMULINK.

From the equations of continuity the pressures p_A and p_B are inferred:

$$\dot{p}_A = [Q_A - A_K \cdot (\dot{x} - \dot{x}_2) - a_{11} \cdot (p_A - p_B)] \cdot \frac{E_{ers}}{V_{0K} + A_K \cdot (x - x_2)} \tag{33.9}$$

$$\dot{p}_B = [A_R . (\dot{x} - \dot{x}_2) + a_{11} \cdot (p_A - p_B) - a_{12} . (p_B - p_0) - Q_B] \cdot \frac{E_{ers}}{(V_{0R} + A_R \cdot h) - A_R \cdot (x - x_2)} \tag{33.10}$$

where a_{11}, a_{12} $[(N/m^2)/(m^3/s)]$, the gradients of leakages; $V_{0K,0R}$ $[m^3]$, the initial average volume of chambers A and B, respectively; E_{ers} $[N/m^2]$, the equivalent bulk modulus of oil; $p_{A,B}$ $[N/m^2]$, the fluid pressure in chambers A and B, respectively; and h [m], the stroke of the piston-rod.

The LHM operation is based on the equations described above, namely (33.7)–(33.10). The MM proposed in this section is implemented in SIMULINK 2.1/ MATLAB 5.1 and has the structure presented in Fig. 33.10.

The signal generator icon from the above figure generates the displacement of the wall x_1, which has a sinusoidal form with the frequency 0.5 Hz and amplitude 0.0001 m.

The subsystems Ssys1 and Ssys2 have as outputs the valve displacement x_v, and the flows Q_A and Q_B, respectively. Ssys3 is the block that implemented Eq. (33.9), while the subsystem Ssys4 modelled Eq. (33.10). The equations that describe the displacement of the wall cylinder and the LHM piston are modelled by subsystem Ssys5. The reference signal is a step one whose values are in the range 0–10 V.

33.4 Conventional Controllers Used to Control the Electrohydraulic Axis

This section is organized as follows: the first part presents the bibliographic research concerning the traditional directions of the control system, and the second one contains the testing of several classic control structures (PID and control algorithms with Luenberger observer) through simulations of the electrohydraulic axis endowed with these controllers.

The testing of the MM is performed with SIMULINK and has a goal of the achievement of reference experimental results in order to perform a comparative study of classical controllers and advanced control structures applied to the electrohydraulic axis.

PID, PI, PD with Filtering

The conventional control structures used in this chapter are PI (proportional-integral), PID (proportional-integral-derivative), and PD (proportional-derivative) with filtering coefficient.

The transfer function of the PI controller has the following expression:

$$H_{\mathrm{PI}} = \frac{U(s)}{\varepsilon(s)} = K_R \cdot \left(1 + \frac{1}{T_i \cdot s}\right) = K_R \cdot \frac{T_i \cdot s + 1}{T_i \cdot s} \tag{33.11}$$

K_R is the proportional factor, and T_i is the time constant of the integrative component.

The transfer function of the PID controller is described by the following equation:

$$H_{\mathrm{PID}} = \frac{U(s)}{\varepsilon(s)} = K_R \cdot \left(1 + \frac{1}{T_i \cdot s} + T_d \cdot s\right) = K_R \cdot \frac{T_i \cdot T_d \cdot s^2 + T_i \cdot s + 1}{T_i \cdot s} \tag{33.12}$$

K_R and T_i have the same significance as previously mentioned, and T_d is the time constant of derivative component.

The transfer function corresponding to PD with filtering has the expression:

$$H_{\mathrm{PDF}}(s) = \frac{U(s)}{\varepsilon(s)} = K_R \cdot \frac{1 + T_d \cdot s}{1 + \alpha \cdot T_d \cdot s} \tag{33.13}$$

where the coefficient α could have values in the range 0.1–0.125.

Generally speaking, PID controllers are commonly used in industrial control systems and, therefore, are well established. Nevertheless, the results obtained using a PID controller for complex control loops are not very satisfactory because it could be costly and time consuming to retune such regulators. PI controller is enough in situations where derivative action is not frequently used.

Observer

The theory of observers, started with the work of Luenberger and Ackermann, is fairly complete and comprehensive. For the proposed axis, an $(n - m - 1)$ order structure of the observer is adopted, where $n = 5$ represents the order of the system and $m = 1$ is the number of outputs [25]. The model of the servodrive is described by five state variables: two of them for the second-order model of the servovalve and the other three for the third-order servoactuator. The use of a linear observer as a parallel model reconstitutes the state-variables of the installation and delivers them to the controller. Two possibilities could be followed: with partial and with global reconstruction. The solution chosen was with partial reconstruction [12]. The complete system consists of the installation with nonlinearities, a parallel second-order model for the servovalve, a third-order linear servoactuator, a correction matrix for the observer, and a controller with five loops for the five state variables [23,25,26].

The block diagram of an electrohydraulic axis controlled with a third-order observer is shown in Fig. 33.11, where $\mathbf{A}$, $\mathbf{b}$, $\mathbf{c}^{\mathrm{T}}$ are the characteristic matrices of the linear system (electrohydraulic axis), $\mathbf{k}$ is the correction matrix, and $\mathbf{R}$ represents the matrix corresponding to the controller.

The simulation diagram of the electrohydraulic axis, controlled by a third-order observer, as it is depicted in SIMULINK, is illustrated in Fig. 33.12.

The algorithm used to compute the matrix **k** and **r** consists of the following steps:

i. achievement of the MM for servovalve and servoactuator;
ii. setting the state-variables of the process;
iii. obtaining the controller upon the dynamics of the closed-loop system;
iv. computing the correction matrix by using the desired poles for the observer [12].

FIGURE 33.11 Block diagram of the control loop using the observer.

FIGURE 33.12 Control structure with observer for the electrohydraulic axis with SIMULINK.

The Linear Mathematical Models (LMM)

The LMM of the servovalve and of the servoactuator are obtained by using the appropriate transfer functions H_{SV}, H_Z:

$$H_{SV}(s) \triangleq \frac{Q_V(s)}{I_V(s)} = \frac{k_V\omega_V^2}{s^2 + 2D_V\omega_V s + \omega_V^2} \tag{33.14}$$

$$H_Z(s) \triangleq \frac{Y_Z(s)}{Q_V(s)} = \frac{k_Z\omega_Z^2}{(s^2 + 2D_Z\omega_Z s + \omega_Z^2)s} \tag{33.15}$$

with the following meanings: Q_V, the servovalve's flow; I_V, the current intensity, Y_Z, the rod's position.

LMM in the State of Space

The used variables are: x_1, the rod position; x_2, the rod velocity; x_3, the rod acceleration; x_4, the spool position; and x_5, the spool velocity.

$$
\begin{aligned}
\dot{x}_1 &= x_2(t), \quad \dot{x}_2 = x_3(t), \quad \dot{x}_4 = x_5(t) \\
\dot{x}_3 &= -\omega_Z^2 x_2(t) - 2D_Z\,\omega_Z x_3(t) + k_Z\,\omega_Z^2 x_4(t) \\
\dot{x}_5 &= -\omega_V^2 x_4(t) - 2D_V\,\omega_V x_5(t) + k_V\,\omega_V^2 u(t)
\end{aligned}
\tag{33.16}
$$

Thus the MM of the axis in state-space form becomes

$$
\begin{aligned}
\dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{b}\mathbf{u}(t) \\
\mathbf{y}(t) &= \mathbf{c}^T\mathbf{x}(t)
\end{aligned}
\tag{33.17}
$$

where

$$
\mathbf{A} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & -\omega_Z^2 & -2D_Z\omega_Z & k_Z & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\omega_V^2 & -2D_V\omega_V \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ k_V \end{pmatrix}, \quad \mathbf{c}^T = (1 \;\; 0 \;\; 0 \;\; 0 \;\; 0) \tag{33.18}
$$

Controller Design

The characteristic polynomial is obtained from $\det[s\mathbf{I} - (\mathbf{A}_C - \mathbf{b}_C\mathbf{r}^T)] = 0$, where $\mathbf{A}_C$ and $\mathbf{b}_C$ are the controllable forms of the matrices $\mathbf{A}$ and $\mathbf{b}$. If $\mathbf{A} \neq \mathbf{A}_C$, the use of transformation matrix $\mathbf{T}$ is advisable, in order to obtain $\mathbf{A}_C$ and $\mathbf{b}_C$. Thus $\mathbf{A}_C = \mathbf{TAT}^{-1}$, and $\mathbf{b}_C = \mathbf{Tb}$. The matrix $\mathbf{F} = \mathbf{A}_C - \mathbf{b}_C\mathbf{r}^T$ has the form

$$
\mathbf{F} = \begin{pmatrix} 0 & 1 & .. & 0 \\ 0 & 0 & .. & 0 \\ . & . & .. & . \\ 0 & 0 & .. & 1 \\ -a_0 - r_1 & -a_1 - r_2 & .. & -a_{n-1} - r_n \end{pmatrix} \tag{33.19}
$$

The characteristic polynomial of the matrix $\mathbf{F}$ is

$$
s^n + (a_{n-1} + r_n)s^{n-1} + \cdots + (a_1 + r_2)s + (a_0 + r_1) \tag{33.20}
$$

The poles chosen for the closed-loop determine the polynomial

$$
s^n + p_{n-1}s^{n-1} + p_{n-2}s^{n-2} + \cdots + p_1 s + p_0 \tag{33.21}
$$

The polynomials (33.20) and (33.21) are identical; therefore, the coefficients of matrix $\mathbf{r}_R^T$ are

$$
r_\nu = p_{\nu-1} - a_{\nu-1}, \quad \nu = 1, \ldots, n
$$

If

$$
\mathbf{A} = \mathbf{Ac}, \quad \mathbf{r}^T = \mathbf{r}_R^T
$$

otherwise

$$\mathbf{r}^T = \mathbf{r}_R^T\mathbf{T} \tag{33.22}$$

Correction Matrix Design

The matrix **F** is $\mathbf{F} = \mathbf{A}^* - \mathbf{K}\mathbf{c}^T$, where $\mathbf{A}^*$ is the matrix of the observer. For **F** the chosen poles are $s_1, s_2, \dots, s_n$, and

$$\det[s\mathbf{I}\text{-}\mathbf{F}] = (s - s_1)(s - s_2)\cdots(s - s_n) \tag{33.23}$$

$$\det[s\mathbf{I}\text{-}\mathbf{F}] = s^n + f_{n-1}s^{n-1} + \cdots + f_1 s + f_0 \tag{33.24}$$

From these two equations, the coefficients $k_1, k_2, \dots, k_n$ are obtained.

In this case, $\mathbf{c}^T = (1\ \ 0\ \ 0)$ and the matrix of the third observer is

$$\mathbf{A}^* = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & -\omega_Z^2 & -2D_Z\,\omega_Z \end{pmatrix} \tag{33.25}$$

The correction matrix influences the transient behavior; the further the poles of **F** from the poles of $\mathbf{A}^*$ the quicker the response.

Simulation Results of Electrohydraulic Axis with Conventional Controllers

Based on the above algorithm in order to determine the correction matrix and the controller matrix, the SIMULINK implementation of the observer involves the following values:

$$r_1 = 19.95854, \quad r_2 = 0.069481, \quad r_3 = -7.06024 \times 10^{-4},$$

$$r_4 = -3.158688 \times 10^2, \quad r_5 = -3.451209 \times 10^{-1}$$

for the controller, and

$$k_1 = 1.67 \times 10^{-2},\ \ k_2 = 3.7028 \times 10^4, \quad k_3 = -6.969698 \times 10^6 \text{ for the correction matrix.}$$

When the reference signal is a step signal with $U = 10$ V, the simulation results are shown in Figs. 33.13 and 33.14

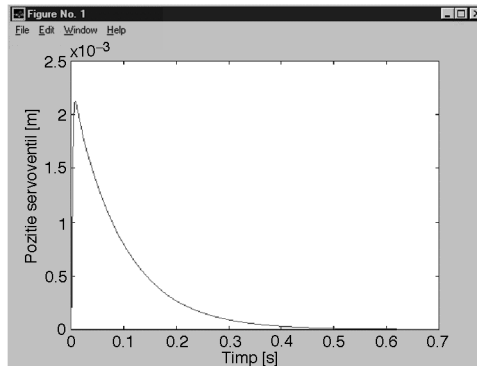



FIGURE 33.13 Position of servovalve for MM with observer.

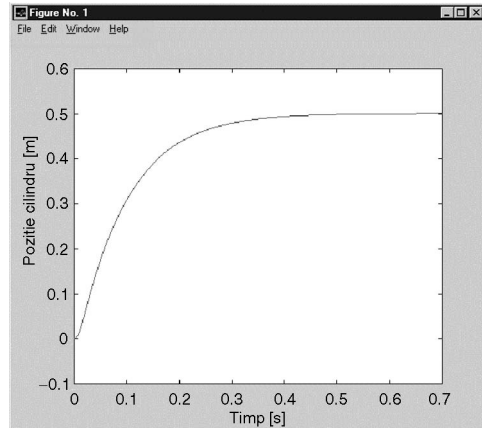



FIGURE 33.14 Cylinder position.

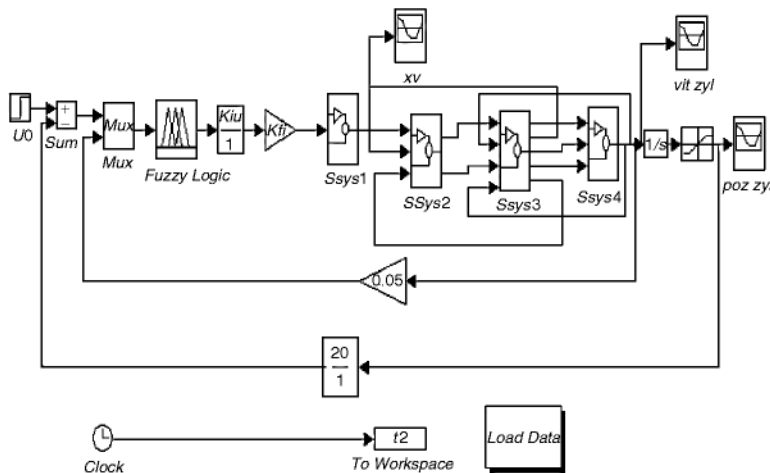



FIGURE 33.15 Electrohydraulic axis control with a fuzzy controller with two inputs.

33.5 Control of Electrohydraulic Axis with Fuzzy Controllers

Section 33.5 of this contribution is devoted to the presentation and the testing of nontraditional controllers based on fuzzy sets, which model the behavior of a human operator in the control process. The simulation results of an electrohydraulic axis with SUGENO and MAMDANI controllers are depicted. For the same number of inference rules extracted from the knowledge base, simulations proved that dynamic performances are improved for a fuzzy controller with two inputs.

The scheme achieved with SIMULINK to control the electrohydraulic axis with two inputs fuzzy controller and a MAMDANI or SUGENO inference is depicted in Fig. 33.15.

The results presented concern the simulation of the hydraulic axis endowed with a fuzzy controller, which is based on MAMDANI inference [59]. "Fuzzy Logic" toolbox gives the user the possibility to create MAMDANI or SUGENO fuzzy systems using graphic interfaces. FIS (Fuzzy Inference System) Editor, Membership Function Editor, and Inference Rules Editor are several of the tools available in SIMULINK. For instance, the corresponding FIS editor and the Membership Function Editor of each input for the proposed fuzzy controller with MAMDANI inference and two inputs are illustrated in Figs. 33.16 and 33.17

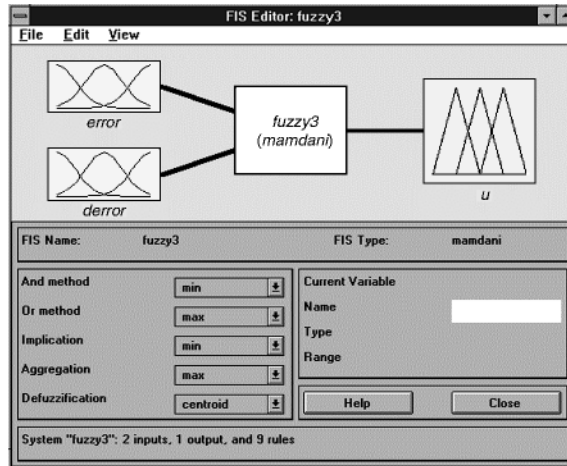



FIGURE 33.16 FIS Editor for fuzzy system based on MAMDANI inference.

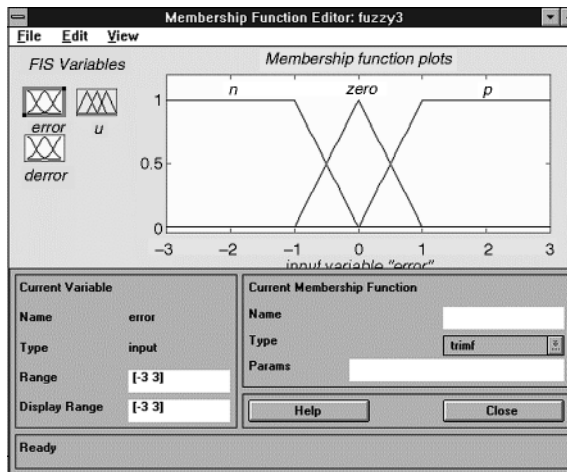



FIGURE 33.17 Membership function associated to the inputs.

For this fuzzy controller there were chosen nine inference rules, which could be visualized using the Inference Rules Editor of SIMULINK. Several simulation results of electrohydraulic axes obtained with the proposed fuzzy controller are shown in Figs. 33.18 and 33.19 and depict graphically the position and the velocity of the cylinder.

33.6 Neural Techniques Used to Control the Electrohydraulic Axis

Section 33.6 has as its goals: to emphasize MATLAB's possibilities of using its resources in order to design control systems based on advanced control techniques such as neural networks; to test through simulation these neural algorithms; and to verify performances of the neural control architecture applied to the studied electrohydraulic axis.

There are two main research directions involved in the neural control. One of these implies the developing of one controller going from a neural network, and the other one embeds several controllers

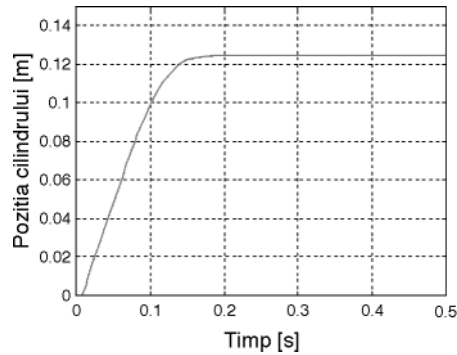



FIGURE 33.18 LHM position.

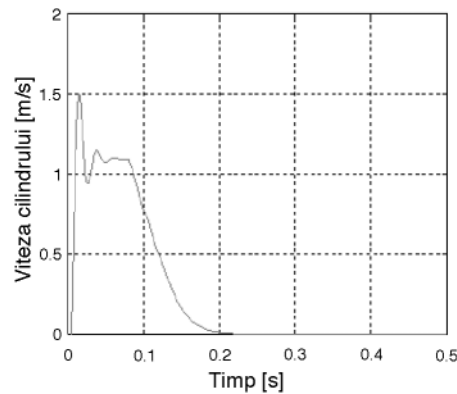



FIGURE 33.19 LHM velocity.

inside a neural network [50]. This section deals with the control of an electrohydraulic axis using a neural controller that has a widely spread structure, namely, multilayer perceptron (MLP).

Neural Control Techniques

Learning Based on Mimic

Inspired from biological systems, learning by mimic is applied to control systems. A supervised neural network can mimic the behavior of another system. A first method to develop a neural controller is to replicate a human controller. The neural controller tries to behave like the human operator. Neural training means learning the correspondence between the information received by the human operator and the control input (Fig. 33.20).

Inverse Learning

The purpose of inverse control is to control a system by using its inverse dynamic. In this case, the neural network receives the output of the system as input and has the input of the system as output. The system works in open loop and has to be in the region where the controller will operate. Inverse learning (Fig. 33.21) is an indirect approach to minimize the network output error instead of the overall system error.

Specialized Inverse Learning

According to Psaltis, who proposed in 1988 a *specialized inverse learning*, the neural network should be trained online in order to minimize the control error $e_y = r - y$ (see Fig. 33.22).

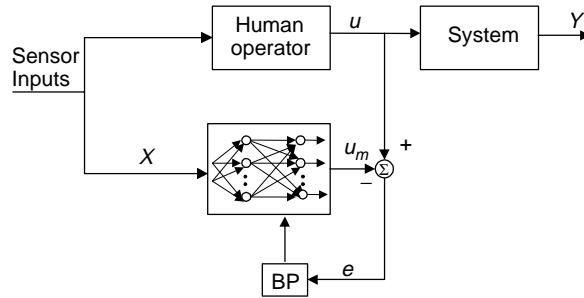



FIGURE 33.20 Diagram for learning based on mimic.

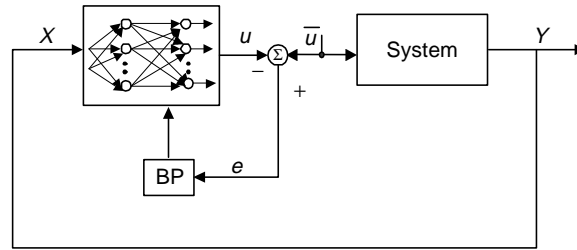



FIGURE 33.21 Training phase at inverse learning.

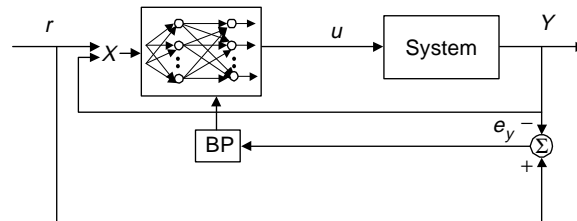
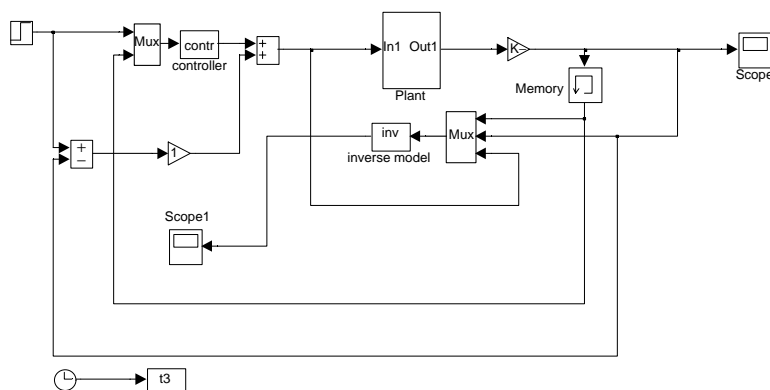



FIGURE 33.22 Specialized inverse control architecture (after [50]).

The neural controller used to control the position of an electrohydraulic axis is a *feed-forward* multi-layer neural network, whose learning algorithm is *back-propagation*. In order to adapt the weights which preserve the learned information, two steps are gone through with: a *forward* propagation procedure of the useful signal and a *backward* propagation of the error. The control structure is implemented in SIMULINK as it is shown in Fig. 33.23. The neural control of the electrohydraulic axis and the achievement of controller parameters are performed online.

A neural network with four layers, having two neurons on the first layer, a neuron on the last layer, and five neurons on each hidden layer, is proposed. The graphic characteristic corresponding to the axis position and obtained using the neural network described above is illustrated in Fig. 33.24.

33.7 Neuro-Fuzzy Techniques Used to Control the Electrohydraulic Axis

This chapter deals with several computer-aided design techniques of hybrid control algorithms. This paper concentrates on these types of algorithms, because the performances achieved through simulation of an electrohydraulic axis with a neuro-fuzzy controller are comparable or superior to those yielded by other control algorithms. Taking into account the novelty of neuro-fuzzy algorithms and the absence in

FIGUGE 33.23 The control structure for proposed controllers.

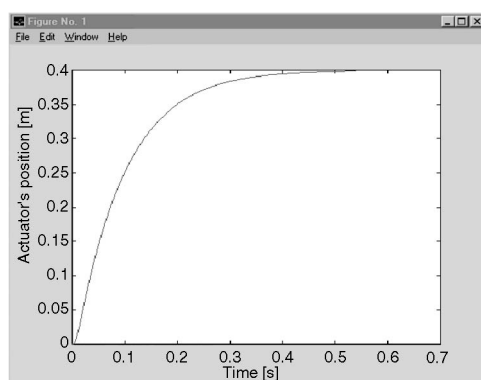



FIGURE 33.24 The axis position for $U = 8$ V input voltage.

SIMULINK of a toolbox devoted to them, the research was oriented to the achievement of a library of C++ programs, which can afford the use of SIMULINK in the design of such controllers. Thus, online adaptation procedures of fuzzy controller parameters are implemented. The comparative study of different classic and advanced algorithms is performed on the basis of integral squared error computed on the transitory horizon.

Because of the capability of fuzzy systems to treat imprecise information, they are strongly recommended in order to express knowledge in the form of linguistic rules. In this way, the human operator's knowledge, which is linguistic or numerical, is used to generate the set of fuzzy if-then rules as a basis for a fuzzy controller. A main drawback of fuzzy systems is the difficulty to design them based on a systematic methodology. To overcome this drawback, the learning procedures from neural networks are successfully applied in order to tune the parameters of membership functions. The merger of neural networks and fuzzy logic has led to the existence of neuro-fuzzy controllers. It can be asserted that neuro-fuzzy controllers embed essential features of both fuzzy systems and neural networks.

The proposed neuro-fuzzy controller has a structure based on the Takagi-Sugeno method and it is depicted in Fig. 33.25.

A learning procedure in fact represents a parameter estimation problem. The learning procedure for the proposed neuro-fuzzy controller is gradient-descendent. The method applied to design such a controller is called inverse learning in which an online technique is used to model the inverse dynamics of the plant. The obtained neuro-fuzzy model—the inverse dynamics of the plant—is used to generate control actions.

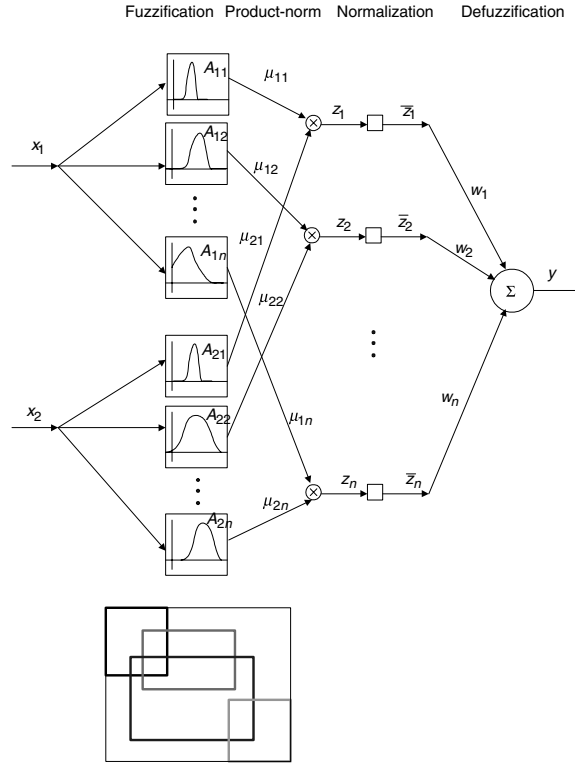



FIGURE 33.25 Structure of the neuro-fuzzy controller.

The neuro-fuzzy controller is a multilayer connectionist system, a multi-input and single-output fuzzy logic system. The network has three layers: one input layer with $n \times m$ units, one hidden layer with n units, and one output layer with one unit [15]. The partition used for this model is a scatter partition [33].

Figure 33.25 presents a particular case where the fuzzy controller has only two inputs and one output. In a general case, the fuzzy controller has m inputs and one output.

The fuzzy rule base contains a set of n linguistic rules in the form:

R_i: If x_1 is A_{1i} and x_2 is A_{2i}
and…
and x_m is A_{mi}
then y is w_i, $i = 1, 2,\ldots, n$

where i is the index of the rule; A_{ji} is a fuzzy set for the jth linguistic variable and the ith rule; and w_i is a number that represents the consequent part.

The membership functions assigned to each input are Gaussian functions. The centers of the membership functions are chosen such that these functions are uniformly distributed over the universe of discourse:

$$\mu_{ji} = e^{-(x_j - a_{ji})^2/2b_{ji}^2} \tag{33.26}$$

The fuzzy inference involved in this neuro-fuzzy controller is the product operator T-norm defined as an *and* conjunction. The firing strength of every rule is

$$z_i = \mu_{1i} \cdot \mu_{2i} \cdot \ldots \cdot \mu_{mi}, \quad i = 1,\ldots,n \tag{33.27}$$

The output is a crisp value obtained as a result of the evaluation of a center of gravity:

$$y = \frac{\sum_{i=1}^{n} z_i \cdot w_i}{\sum_{i=1}^{n} z_i} = \frac{z_1}{\sum_{i=1}^{n} z_i} \cdot w_1 + \frac{z_2}{\sum_{i=1}^{n} z_i} \cdot w_2 + \cdots + \frac{z_n}{\sum_{i=1}^{n} z_i} \cdot w_n = \sum_{i=1}^{n} \bar{z}_i \cdot w_i \tag{33.28}$$

The parameters to be estimated are obtained by finding the minimum of the following cost function:

$$J(k) = \frac{1}{2} \cdot (y(k) - y_d(k))^2 \tag{33.29}$$

where $y_d(k)$ is the desired output and $y(k)$ is the obtained response at time k.

To minimize this cost function, the stochastic approximation method is used. The learning procedure means the estimation of parameters and is based on the least-mean square algorithm. The parameters to be estimated are

$$p = (a_{11}, \ldots, a_{nm}, b_{11}, \ldots, b_{nm}, w_1, \ldots, w_n) \tag{33.30}$$

The equations to adapt the parameters are the following:

$$\begin{aligned} a_{ji}(t+1) &= a_{ji}(t) - \boldsymbol{\lambda}_a \frac{z_i}{\sum_{l=1}^{n} z_1} \cdot (y - y_d) \cdot (w_i - y) \cdot \frac{x_j - a_{ji}(t)}{b_{ji}^2} \\ b_{ji}(t+1) &= b_{ji}(t) - \boldsymbol{\lambda}_b \frac{z_i}{\sum_{l=1}^{n} z_1} \cdot (y - y_d) \cdot (w_i - y) \cdot \frac{(x_j - a_{ji}(t))^2}{b_{ji}^3} \\ w_i(t+1) &= w_i(t) - \boldsymbol{\lambda}_w \frac{z_i}{\sum_{l=1}^{n} z_1} \cdot (y - y_d) \end{aligned} \tag{33.31}$$

where the learning factors λ_a, λ_b, λ_w are predefined.

In the learning process, parameters that could be modified are (a_{ji}, b_{ji}) which describe Gaussian functions, and w_i, the conclusion values. If the structure of the membership function is established, the only values that could be modified are w_i.

Control Structure

In order to design the neuro-fuzzy controller proposed above, the inverse learning method is applied. The control of an electrohydraulic axis involves the use of an online technique to model the inverse dynamics of the plant. The block diagram for online inverse learning is presented in Fig. 33.26.

This scheme is in open loop and it is also found by the Controller Output Error Method (COEM) [1] to online tune or adapt the parameters of a fuzzy controller. This method does not require the plant output error to be propagated at the input. There is another constraint, namely the controller has to be capable of stabilizing the plant before the commencement of tuning. To avoid this requirement, a modified COEM (MCOEM) [2] is used. The diagram block in this case is depicted in Fig. 33.27.

A proportional feedback controller P is introduced and in this situation the plant input is the sum of $u'(k)$ and $u_p(k)$. The consequent singletons are initialized to zero and the controller P is chosen in such a way that it stabilizes the plant. The structure and the parameters of inverse model and of neuro-fuzzy controller are identical.

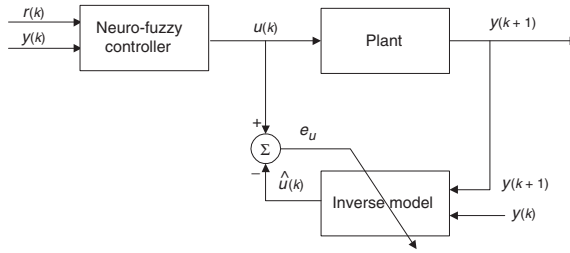



FIGURE 33.26 Diagram of control based on inverse learning.

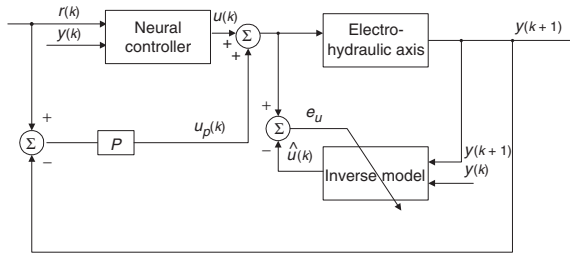



FIGURE 33.27 Block diagram for inverse learning with proportional controller.

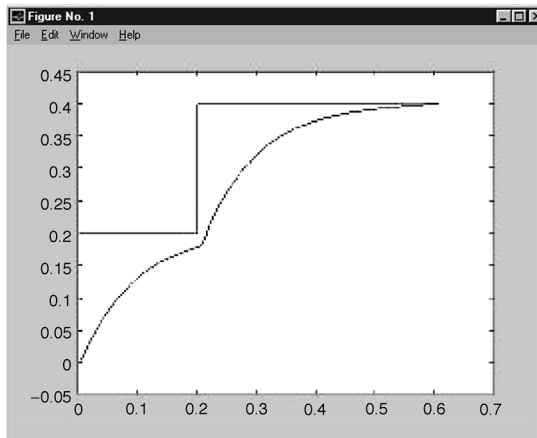



FIGURE 33.28 The position control with neuro-fuzzy controller.

There are two phases in the design of such a controller: the control and the adaptation. In the control phase, the plant output and the reference signal determine a control command $u(k)$. The plant input becomes $u(k)$, the sum of the $u(k)$ and $u_p(k)$. In the adaptation phase, the inverse model, which has as inputs $y(k + 1)$ and $y(k)$, produces the signal $\hat{u}(k)$ as an output. This signal is used to compute the error $e_u(k)$, which determines the value of the cost function $J(k)$ that has to be minimized.

$$J(k) = \frac{1}{2} \cdot e_u^2(k) = \frac{1}{2} \cdot (u(k) - \hat{u}(k))^2 \qquad (33.32)$$

This procedure was used at the control of the electrohydraulic axis position, where the controller parameters are determined online. The actuator position obtained when the reference signal is changed from $U = 4$ V to $U = 8$ V is depicted in Fig. 33.28.

TABLE 33.1

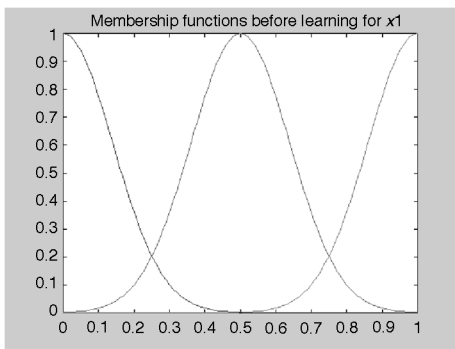
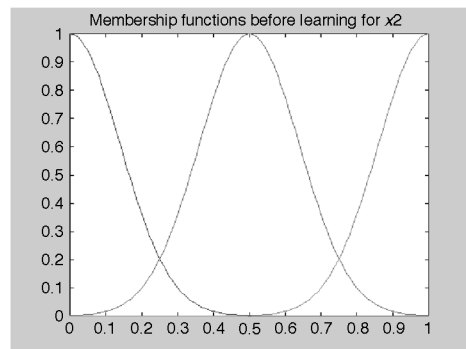
Regulator	IAE	ISE
PID (chapter 4)	0.8042	3.4754
PI (chapter 4)	0.8006	3.4618
PD (chapter 4)	0.7928	3.4537
Neural (chapter 6)	0.8027	3.4622
Neuro-fuzzy (chapter 7)	0.7911	3.4501

(a)

(b)

FIGURE 33.29 (a) Membership functions before learning for the variable x_1, (b) Membership functions before learning for the variable x_2.

In order to achieve a comparison of the modern control algorithms (included in this thesis) to the conventional structures, two spread integral criteria, namely, the integral of absolute error (IAE) performance index and the integral of squared error (ISE), are used. The results obtained applying these criteria are included in Table 33.1.

According to previous results, it can be inferred that the described neuro-fuzzy controller exhibits superior performances compared to those obtained with the neural controller based on MLP, or with the classic controllers (PID, PI, PD with filtering) presented in this paper. The simulation results emphasize the neuro-fuzzy controller, arguing that it represents a very useful tool for practical applications with many nonlinearities.

Optimized results were obtained through variation of data sets and number of iterations. In order to test the performance of the proposed neuro-fuzzy controller, one nonlinear function given by an analytical equation was approximated. The membership functions of input variables x_1 and x_2 before learning are shown in Figs. 33.29(a,b). The surface obtained after simulation is depicted in Fig. 33.30(c). One may observe the accuracy of the reconstruction after 300 learning iterations by comparison with the surface to be obtained.

Sets of intermediary results obtained with different simulation data sets are presented below. Different data sets of simulations were used in order to achieve optimized results. Some of them are presented in Figs. 33.30–33.34 without comment.

In order to obtain good performances from the model, 10 membership functions are used for each input variable. The learning factors λ_a, λ_b, λ_w were chosen as 0.01. The control algorithm is capable of handling the change in operating range. The results of the electrohydraulic axis simulation with the proposed neuro-fuzzy controller are obtained for various inputs. Those in time domain, results presented in Figs. 33.35(a,b), correspond to input voltages of 8 and 10 V.

1st set: gw = 0.1; ga = 0.05; gb = 0.05; nepoc = 100; nesant = 100; niter = 200; threshold error = 0.001; Vmax = 0.04.

(a)

(b)

(c)

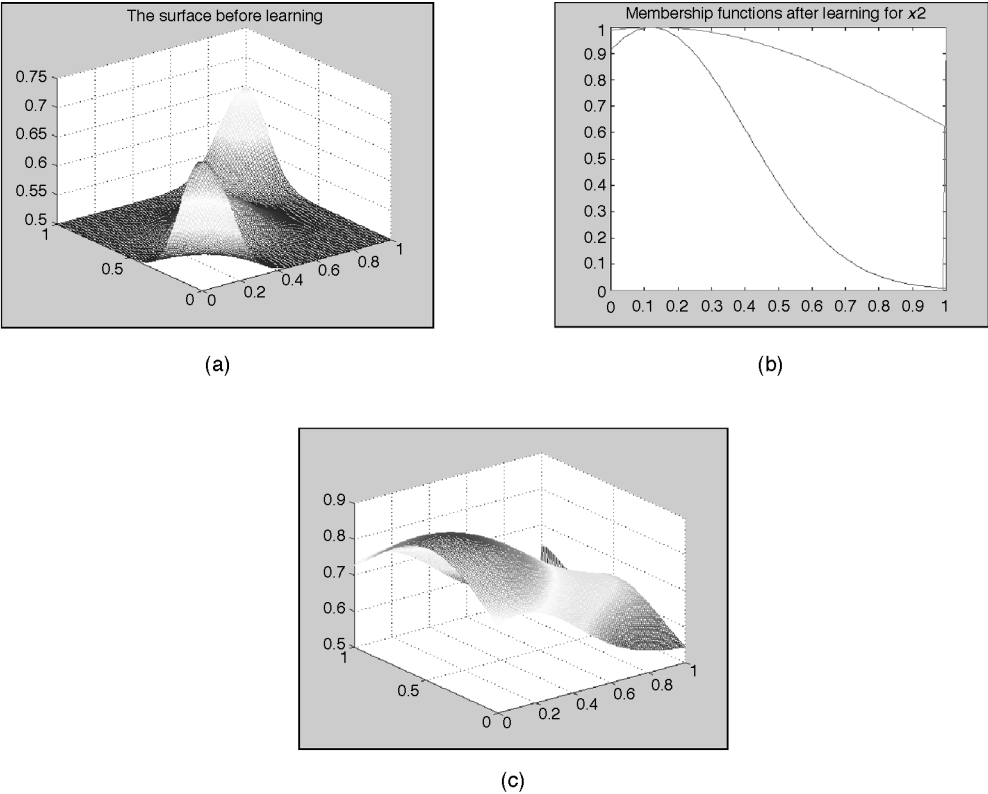



FIGURE 33.30 (a) The surface obtained after the first iteration, (b) Membership functions after learning for the variable x_2, (c) The surface obtained after simulation.

2nd set: gw = 0.1; ga = 0.07; gb = 0.05; nepoc = 100; nesant = 21; niter = 200; threshold error = 0.001; Vmax = 0.0475.

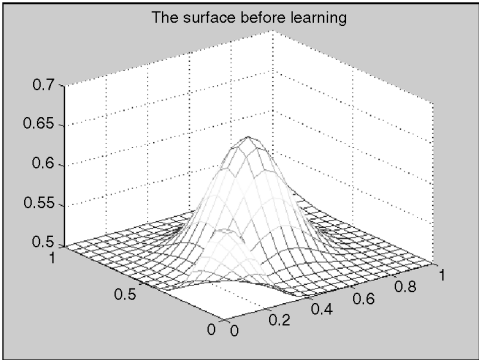



FIGURE 33.31 The surface obtained after the first iteration.

3rd set: gw = 0.5; ga = 0.07; gb = 0.03; nepoc = 200; nesant = 21; niter = 200; threshold error = 0.001; Vmax = 0.047515.

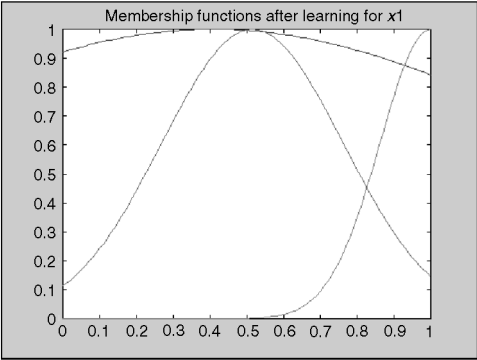



FIGURE 33.32 Membership functions after learning for the variable x_2.

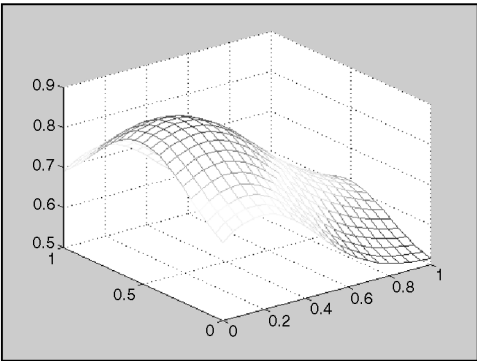



FIGURE 33.33 The surface obtained after learning.

33.8 Software Considerations

The MM of electrohydraulic axis studied in this thesis is supported by a physical installation existing in the mechatronics laboratory of UAS-Konstanz (see Fig. 33.1b). Two variants of nonlinear MM are set forth in Section 33.3 and add in static and dynamic nonlinearities that arise in the function of electrohydraulic axis [23, 58].

The MM of hydraulic drive presented in the structure of ROBI_3 was implemented in SIMULINK in order to study the dynamic behavior of the axis [26, 27]. The extended variant of MM hydraulic axis was done taking into account the relative motion of the constituent parts of this servodrive.

The neural and neuro-fuzzy controller (Takagi-Sugeno) was developed in Borland C++ and implemented in SIMULINK for controlling the electrohydraulic axis. SIMULINK offers the user a FUZZY LOGIC library that allows the designing and modeling of SUGENO or MAMDANI fuzzy inference systems. The lack of a dedicated software to design neuro-fuzzy controllers persuaded the implementation of such a controller in C++ and afterwards the use of it in SIMULINK [26,27,28].

The support for simulation, SIMULINK 2.1 and MATLAB 5.2 (under Windows), offers solutions to implement our controllers as modules and corresponding icon in a specialized toolbox. In our experiments, we used the facility offered by S-functions and C MEX in conjunction with Borland C++ 5.0 to compile them. We have chosen the C S-function because of the speed necessary to process the information in our block that implement the controller. The block that implements the controller has two inputs (even three inputs are available, though the adaptation process is more complicated) and one output.

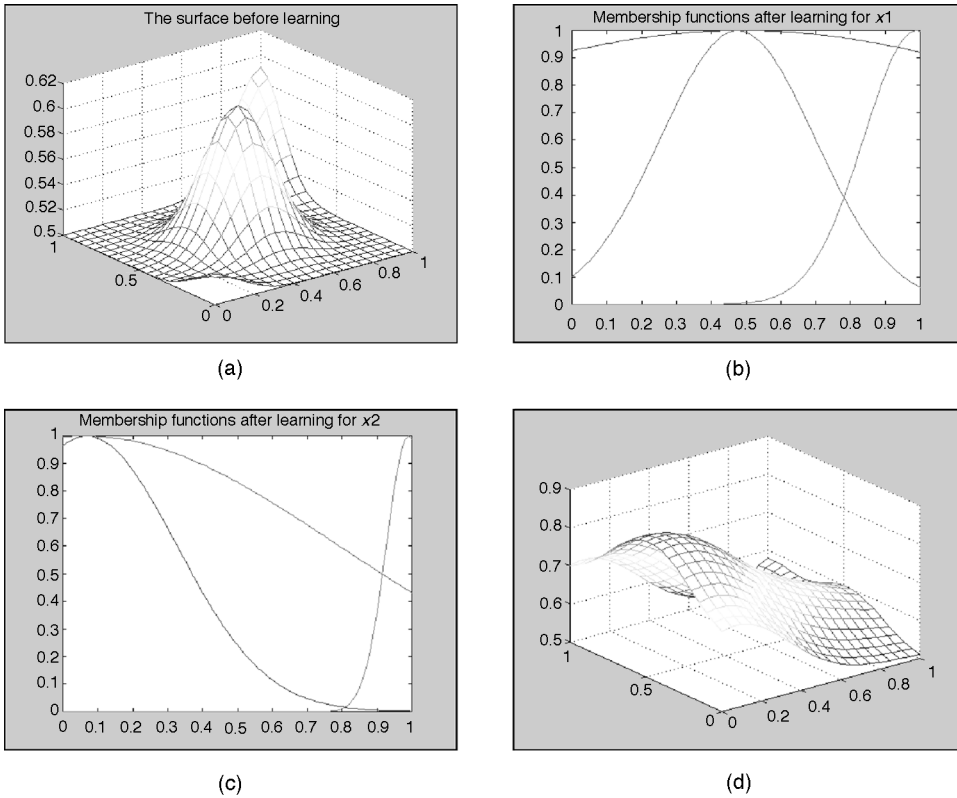



FIGURE 33.34 (a) The surface obtained after first iteration, (b) Membership functions after learning for the variable x_1, (c) Membership functions after learning for the variable x_2, (d) The surface after learning.

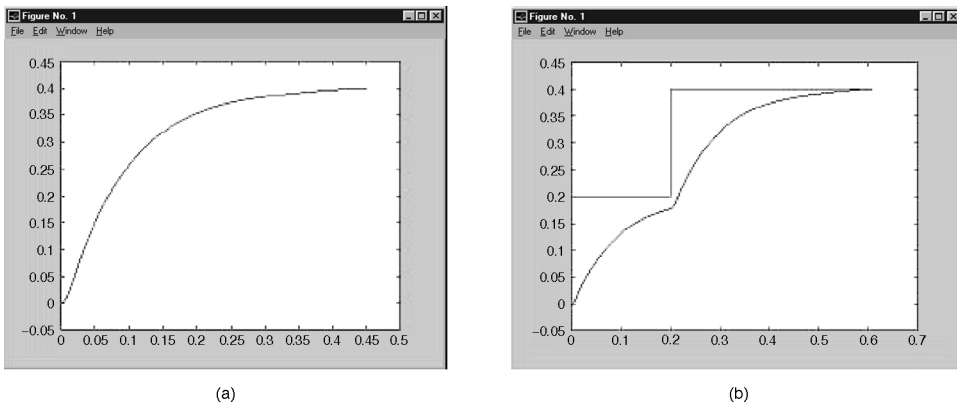



FIGURE 33.35 (a) The position control with neuro-fuzzy controller (U = 8 V), (b) The position control with neuro-fuzzy controller (U = 10 V).

The adapting parameters (weights, centers, and spread for Gaussian function) must be persistent. Declaring global, static or using the workspace in order to store the are useful techniques to accomplish the task.

The newest version of Simulink offers the possibility to write wrapper S-function, to use the callbacks functions, and as an alternative ADA or Fortran programming language.

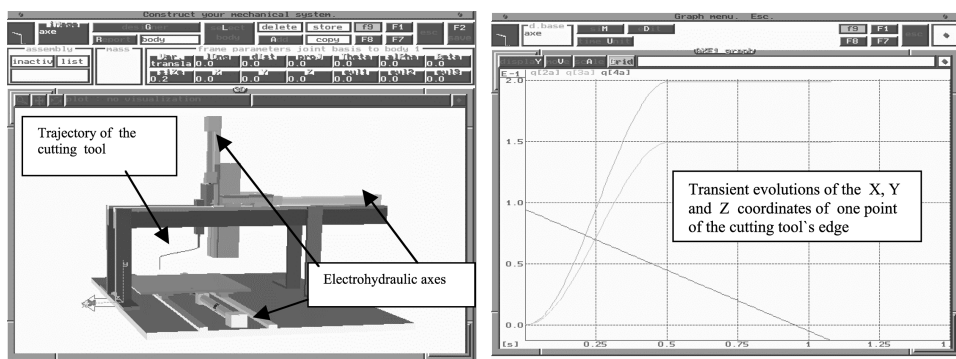



FIGURE 33.36 3-Axes Cartesian robot modeled, controlled, and simulated in SDS [21,24].

33.9 Conclusions

The research achieved as part of this chapter has, as an essential purpose, the development of improved control structure based on advanced techniques (neural and fuzzy), in relation to the conventional one. The studied electrohydraulic axis is a component of the Cartesian robot ROBI_3 implemented in the mechatronics laboratory of UAS-Constance. 3D-simulations with direct and inverse dynamics and implemented controllers by using the SDS-Modelling and Simulation software of the real installation were performed [21,23,25,28]. Model and simulation results are presented in Fig. 33.36.

As an overview, Section 33.1 deals with the introduction of this chapter. In Section 33.2, the most important aspects of electrohydraulic system control and of nonlinearities that arise with this type of installation are pointed out. The robot, ROBI_3, is presented from both a component and control perspective.

The mathematical model (MM) of ROBI_3's hydraulic axis is described in Section 33.3. The nonlinear MM is achieved based upon technical data of different components of installation and also taking into account theoretic assessments of electrohydraulic installation functionality. Simulation results of nonlinear MM placed into a position loop are obtained with simulation environment SIMULINK/MATLAB. As a general remark, it should be mentioned that all simulations included in this chapter are achieved by using MATLAB/SIMULINK, while the advanced control algorithms (neural and neuro-fuzzy) are developed in Borland C++ 5.0.

Section 33.4 contains a short overview of the theory devoted to conventional controllers PID, PD, PI, and observer, followed by simulation results of the electrohydraulic axis endowed with the above control structures.

Following the scientific goal of this contribution, Section 33.5 reviews fuzzy system theory and presents simulations of electrohydraulic axis with fuzzy controllers. Fuzzy system theory has contributed greatly to system modeling, and the development of a theoretical frame appropriate to implement the qualitative reasoning specific to human beings. This kind of reasoning is very useful to model complex systems, which are characterized by nonlinearities or imprecise information. Simulation results of hydraulic axis are obtained using SUGENO and MAMDANI fuzzy controllers.

A short introduction of neural networks theory, the most widespread neural structures and also neural control techniques are presented in the beginning of Section 33.6. Neural networks work quantitatively, numerically. If fuzzy logic has an inference based on uncertainty, then neural networks learn by training, at the end of which the network will approximate a desired function. The analysis of trained NN involves many challenges, and as a result, the rules are usually not extracted from trained NN. Simulation results included in Section 33.6 are obtained with a multilayer NN.

Neuro-fuzzy systems preserve the characteristics of NN and also of fuzzy systems, and have been used successfully in control in recent years. Section 33.7 is devoted to the neuro-fuzzy system theory, to the presentation of neuro-fuzzy controller implemented in Borland C++ and applied in SIMULINK to

electrohydraulic axis, to the simulation results achieved in this case, and to the comparative study of conventional and modern controllers.

Section 33.8 contains a concise presentation of this chapter, the main contributions to the subject area presented, as well as a listing of perspective areas of interest in order to pursue further research in this direction.

Without intending to confine the parameters of this chapter, following is a listing of possible research directions and development perspectives that may be followed in future research endeavors:

- applying various controllers implemented in SIMULINK not only to control the electrohydraulic axis discussed, but also for systems with very complex structure which are involved in large hydraulic installations, offering the user a neuro-fuzzy controller's library;
- the hardware implementation of described neuro-fuzzy controller;
- continued research in the development of an optimal controller, systemically based (through the further study of stability utilizing linear matrix inequalities—LMI);
- the integration of presented controllers in software packages dedicated to hydraulic and pneumatic fields (for instance in HYPAS[23], DSH, etc.);
- the development of controller design in order to promote those controllers, which allow a better symbiosis between classical and advanced methods (neuro-fuzzy, genetic algorithms);
- the extension of preoccupations and extrapolation of research results regarding control of velocity, acceleration, pressure, flow, force, moment, and power.

References

1. Andersen, H.C., Lotfi, A., Tsoi, A.C. A new approach to adaptive fuzzy control: the controller output error method, *IEEE Trans. on Systems, Man, and Cybernetics,* SMC-27-B(4), August 1997.
2. Abonyi, J., Nagy, L., Szeifert, F. Indirect adaptive Sugeno fuzzy control, *Proceedings in Artificial Intelligence,* FNS'98, München, Germany,19–20 martie.
3. Backé, W. *Systematik der hydraulischen Widerstandsschaltungen in Ventilen und Regelkreisen*. Krauskopf-Verlag, Mainz, 1974.
4. Costa Branco, P.J., Dente, J.A. *Inverse-Model Compensation Using Fuzzy Modeling and Fuzzy Learning Schemes.* Intelligent Engineering Systems through Artificial Neural Networks, Smart Engineering Systems: Fuzzy Logic and Evolutionary Programming, Ed. C.H. Dagli, M. Akay et al. Vol. 6, ASME Press, New York, pp. 237–242, 1996.
5. Brown, M., Harris, C. *Neuro-fuzzy Adaptive Modelling and Control,* Prentice-Hall, Englewood Cliffs, NJ, 1994.
6. Catana, I., Vasiliu, D., Vasiliu, N. *Servomecanisme electrohidraulice*. *Constructie, functionare, modelare, simulare si proiectare asistata de calculator*. U.P.B. Bucuresti, 1995.
7. Cybenko, G. *Mathematical Problems in Neural Computing.* Signal Processing Scattering and Operator Theory and Numerical Processing, Vol. 3, Kashoek, M.A., van Schupper, J.H., Ram, A.C. Ed., 1989, pp. 47–64.
8. Driankov, D., Hellendoorn, H., Reinfrank, M. *An Introduction to Fuzzy Control.* Springer-Verlag, Berlin, 1993.
9. Dubois, D., Prade, H., Ughetto, L. Checking the coherence and redundancy of fuzzy knowledge bases, *IEEE Trans. on Fuzzy Systems,* 5(5):398–417, 1997.
10. Dumitrache, I. sa. *Automatizari electronice.* Editura Didactica si Pedagogica, Bucuresti, 1993.
11. Dumitrache, I., Catana, I., Militaru, A. *Fuzzy Controller for Hydraulic Servosystems.* IFAC International Workshop on Trends in H& P Components & Systems, Chicago, IL, 1994.
12. Föllinger, O. *Regelungstechnick*. Dr. A. Hütig Verlag, Heidelberg, Germany, 1978.
13. Friedrich, A. *Logik und Fuzzy-Logik*. Expert-Verlag,1997.

14. Ghaoui, L. El. Reduced-order multimodel control using linear matrix inequalities: sufficient conditions, *Proc. Od ACC 1993,* 1993, pp. 633–634.
15. Godjevac, J., Steele, N. Adaptive neuro-fuzzy controller for navigation of mobile robot, *International Symposium on Neuro-Fuzzy Systems AT'96,* Conf. Report, EPFL-Lausanne, 1996.
16. Gupta, M.M. *Fuzzy Logic and Neural Networks,* Proc. of the 2nd International Conference on Fuzzy Logic & Neural networks, Iizuka, Japan, 17–22 July, 1992, pp. 187–188.
17. Healey, M. *Principles of Automatic Control.* The English Universities Press Ltd., 1975.
18. Haykin, S. *Neural Networks,* MacMillan College Publishing Company, New York, 1994.
19. Ionescu, Fl. Computer aided design of hydraulic and electrohydraulic drive installations, *Proceed. 9th Triennial World IFAC Congress,* Budapesta, Ungaria, Pergamon Press, Vol. 1, 1984, pp. 569–574.
20. Ionescu, Fl., Stoffel, B. *Contribution to the Automatic Generation of Mathematical Models for the Computer Assisted Analysis and Synthesis of Hydraulic Drive Systems.* Proceed. of the 2nd Intern. Conf. on Fluid Power, 19–21 March 1991, Tampere, Finland, pp. 469–482.
21. Ionescu, Fl., Haszler, Fl. *TORCH: A Control Software for Electrohydraulic Cartesians Robots.* Proceed. of the 6th Intern. IMEKO Symposium on Measurement and Control in Robotics, ICMR'96, Bruxelles, Belgium, 9–11 May, 1996, pp. 484–489.
22. Ionescu, Fl. *Non-Linear Problems in the Hydraulic Drive Systems.* 2nd World Congress of Nonlinear Analysts, Athena, Greece, 10–17 July 1996, Pergamon Press, Vol. 30, part 3, pp. 1447–1461.
23. Ionescu, Fl., Vlad, C.I. Tools of HYPAS for the control of electrohydraulic drive installations, *Proc. of 7th Symposium on Computer Aided Control Systems Design,* Gent, Belgia, 1997, pp. 311–316.
24. Ionescu, Fl., Borangiu, Th., Vlad, C.I. High integrated CAD strategies for control design of electro-hydraulic systems, *Proc. 3rd IFAC Conference SSC,* Bucharest, 1997, pp. 390–395.
25. Ionescu, Fl., Vlad, C.I. *Hypas tools for the control of electro-hydraulic drive installations Journal **a**,* Vol. 38, No. 3, Belgium, 1997, pp. 38–41.
26. Ionescu, Fl., Vlad, C.I. Sugeno and hypas fuzzy-control solutions for electro-hydraulic drive installations, *Proceedings EUFIT'97,* Aachen, Germany, 8–11 Sept., Vol. 2, 1997, pp. 1238–1242.
27. Ionescu, Fl., Vlad, C.I., Arotaritei, D. Fuzzy and neuro-fuzzy HYPAS controllers implemented for an electro-hydraulic axis, *International ICSC Symposium on Engineering of Intelligent Systems EIS'98,* Teneriffe, Spain, Feb. 11–13, 1998.
28. Ionescu, Fl., Arotaritei, D., Vlad, C.I. *Modelling of Nonlinearities, Signal Reconstruction and Predictive Solutions Applied in Mechatronics Systems by Using Neuro-Fuzzy Systems,* Internal Report, Department of Mechatronics, FH-University of Applied Sciences-Konstanz, 1998.
29. Ionescu,V., Varga, A. *Teoria Sistemelor. Sinteza robusta. Metode numerice de calcul.* Editura ALL, Bucuresti, 1995.
30. Isermann, R. *Digitale Regelsysteme,* Springer-Verlag, Berlin, 1987.
31. Isermann, R. *Zur Anwendung der Fuzzy-Logik in der Regelungstechnik.* Automatisierungstechnische Praxis (atp) Fuzzy-Control, 38, Oldenbourg Verlag, Germany, 1996.
32. Isermann, R. On fuzzy logic application for automatic control, supervision, and fault diagnosis, *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans,* Vol. 28, No. 2, March 1998, pp. 221–235.
33. Jang, J.-S.R., Sun, C.-T., Mizutani, E. Neuro-fuzzy and soft computing, *A Computational Approach to Learning and Machine Intelligence,* Prentice-Hall, Englewood Cliffs, NJ, 1997.
34. Joh, J., Hong, S.K., Nam, Y., Chung, W.J. On the systematic design of Takagi-Sugeno fuzzy control systems, *International ICSC Symposium on Engineering of Intelligent Systems EIS'98,* Tenerife, Feb. 1998.
35. Kandel, E.R. *Nerves Cell and Behavior,* Principles of Neural Sciences, 3rd ed., 1992, pp. 18–36.
36. Knappe, H. Comparison of conventional and fuzzy-control of non-linear systems, In: Kruse, R., *Fuzzy Systems in Computer Science,* Verlag Vieweg, Wiesbaden, Germany, 1994.
37. Kokotovic, P.V. *Lectures Notes in Control and Information Sciences.* Springer-Verlag, Berlin, 1991.
38. Kosko, B. *Neural Networks and Fuzzy Systems.* Prentice-Hall, Englewood Cliffs, NJ, 1992.

39. Kovacic, Z., Balenovic, M., Bogdan, S. Sensitivity-based self-learning fuzzy logic for a servo-system, *IEEE Control Systems,* June, 1998.
40. Lippmman, R. An introduction to computing with neural nets, *IEEE ASSPMagazine,* April 1987, pp. 4–22.
41. MATLAB 5.2. MathWorks Corp, USA. 1998.
42. Miller, Th., Sutton, R., Werbos, P.J. *Neural Networks for Control,* MIT Press, 1990.
43. Nauck, D., Klawonn, F., Kruse, R. *Neuronale Netze und Fuzzy-Systeme.* Grundlagen des Konnektionismus, Neuronaler Fuzzy-Systeme und der Kopplung mit wissensbasierten Methoden, Vieweg, 1994, Germany.
44. Nesterov, Y., Nemirovski, A. *Interior Point Polynomial Methods in Convex Programming: Theory and Applications,* SIAM, Philadelphia, 1994.
45. Pedrycz, W. *Fuzzy Control and Fuzzy Systems,* Wiley, New York, 2nd ed., 1993.
46. Piechnik, M., Feuser, A. *Simulation mit Komfort - HYVOS 4.0 und MOSIHS 1.0,* Ö & P, 38, 1994.
47. Postlethwaite, B.E. A model-based fuzzy controller, *Trans IChemE,* Vol. 72, Part A, Jan. 1994.
48. Postlethwaite, B.E. Building a model-based fuzzy controller, *Fuzzy Sets and Systems,* 79(1996), Elsevier.
49. Rehfeldt, K., Shöne, A., Büngener, N. *Einsatz von Fuzzy-Reglern zur Drehzahlregelung einer Hydraulikpumpe, Ölhydraulik und Pneumatic,* 36, Nr. 6, pp. 397–402, 1992.
50. Ronco, E., Gawthrop, P.J. *Neural Networks for Modelling and Control.* Technical Report: csc97008, Centre for System and Control, Dept. of Mechanical Engineering, Univ. of Glasgow, 10 Nov. 1997.
51. Simulink, Dynamic System Simulation for MATLAB, Writing S-functions, The Math Works Inc., 1998.
52. Sontag, E.D. Mathematical control theory, *Deterministic Finite Dimensional Systems.* Springer-Verlag, Berlin, 1990.
53. Takagi, T., Sugeno, M. Fuzzy identification of systems and its applications to modeling and control, *IEEE Trans. Systems, Man, and Cybernetics,* Vol SMC-15, No. 1, pp. 116–132, 1985.
54. Tanaka, K., Sugeno, M. Stability analysis and design of fuzzy control systems, *Fuzzy Sets and Systems,* Vol. 45, 1992, pp. 135–156.
55. Teodorescu, H.N. *Sisteme Fuzzy si Aplicatii.* Institutul Politehnic Iasi, Romania,1989.
56. Tertisco, M., Penescu, C., Ionescu, G., Ceanga, E. *Identificarea Experimentala a Proceselor Automatizate.* Editura Tehnica, Bucuresti, 1971.
57. Viersma, T. J. *Analysis, Synthesis and Design of Hydraulic Servosystems and Pipelines.* Elsevier, Amsterdam-New York, 1980.
58. Vlad, C.I. Contributions to the Direct Computer Control of Electrohydraulic Axes for Industrial Robots. Technical University "Politehnica", Bucharest, Romania, 1998.
59. Wang, L., Liu, G.P., Harris, C.J., Brown, M. *Advanced Adaptive Control,* Pergamon, 1997.
60. Werbos, B. *Overview* of *Design and Capabilities.* In Neural Networks for Control, pp. 59–65, MIT Press, MA, 1990.
61. Westcott, J.H. The minimum-moment-of-error-squared criterion: a new performance criterion for servo mechanisms, *Proc. of IEE.,* Measurements Section, pp. 471–480, 1954.
62. Yager, R., Zadeh, L. *Fuzzy Sets, Neural Networks and Soft Computing,* 1994.
63. Zadeh, Lotfi. *Fuzzy sets, Information & Control,* No. 8, pp. 338–353, 1965.
64. Zadeh, L., King-Sun Fu, Tanaka, K., Shimura, M. *Fuzzy Sets and their Applications to Cognitive and Decision Processes.* Academic Press, 1975.
65. Zimmermann, H.-J. *Fuzzy Sets Theory - and Its Applications,* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.

34

Design Optimization of Mechatronic Systems

Tomas Brezina
Technical University of Brno

Ctirad Kratochvil
Technical University of Brno

Cestmir Ondrusek
Technical University of Brno

34.1 Introduction

Electromechanical systems form an integral part of mechanical and mechatronic systems. Their optimization is a necessary condition for a product to be competitive. In engineering practice, a large number of optimization and identification problems exist that could not be solved without the use of computers [5]. The present level of technological development is characterized by increasing the performance of machines with the production costs kept at a satisfactory level. The demands on the reliability and safety of operation of the designed machines are also considerable.

From practical experience we know that the dynamic properties of electromechanical systems have a considerable influence on their reliability and safety. On the other hand, the tendency to push the price of a machine down often leads to unfavorable dynamic properties that result in increased vibrations and noise during operation. Also, electrical properties dramatically deteriorate as the amount of active materials in a machine is reduced. The increased load leads to, among other things, excessive heat formation, which, in turn, has a negative effect on insulation, shortening the service life of a machine.

34.2 Optimization Methods

Principles of Optimization

The properties of electromechanical systems can be described mathematically using physical quantities. The degree of these properties is then described using mathematically formulated objective (preference) functions. Structural parameters ranging between limit values given as satisfying secondary conditions are the independent variables of these functions. The particular form of the functions depends on the type of machine and its mathematical description. The solutions of a mathematically formulated optimization

problem together with optimization methods allow a considerable number of different design variants of a machine to be calculated in a relatively short time. They also make it possible to perform these calculations at production planning stages for a prototype to possess the qualities given by a chosen criteria function. In this way, the design of a machine is not only analyzed but also modified and reconstructed in terms of its electromechanical properties with the aim to improve these properties as much as possible (or optimize them).

From a physical point of view, these are actually problems that, to a certain degree, are inverse to those of calculus. A problem of calculus assumes a fixed, mathematically described model of a real machine to be used for deriving its resulting properties. In problems of calculus, we define properties and try to find out which parameters of a chosen mathematical model possess those properties. In problems of parametric optimization, we look for those parameters that, by a chosen preference function, provide the best properties. It is clear that problems of synthesis and optimization are much more sophisticated than those of calculus.

Parametric Optimization

As the aim is to find the values of certain structural parameters of a machine, we shall deal with this notion in more detail. By a parametric optimization of electromechanical systems, we mean the process of finding those parameters of a mathematical description of the system (arranged in a vector $\bar{p}$) from a set P of admissible parameters at which a suitably selected objective (preference) function $\psi(p)$ of these parameters reaches its extreme.

The objective function $\psi(p)$ quantifies the degree of the properties of an electromagnetic system that has to be made extreme (the parameters with the best degree of this property have to be chosen). When defining an admissible set P, we are guided by the structural possibilities of changes in individual parameters (variables), or we can introduce secondary criteria of the type "the degree of properties may not exceed given critical limits." The possibility of taking into consideration the structural changes of parameters leads to the so-called trivial (natural) constraints of the type $p_i^d \leq p_i \leq p_i^h$, where p_i^d is the lower and p_i^h the upper bound of the ith optimization variable. The introduction of secondary criteria leads to the definition of limiting functions q_i of optimization variables for which we have q_i^d £ $q_i(p)$ £ q_i^h, where q_i^d is the lower and q_i^h is the upper bound of the relevant function.

Thus, from the mathematical point of view, parametric optimization of electromechanical systems is formulated as the problem of finding a point p in the admissible set P, at which the preference function ψ reaches its global extreme value (maximum or minimum) with regard to P. The admissible set is generally described by m inequalities defined by functions $q_j(p)$, where $j = 1, 2,..., m$ If $P = R^s$, where s is the number of variables to be optimized, we say that the optimization is unconditional. In all other cases we say that the optimization is conditional.

To solve the problem of optimizing the selected properties of a system, the following has to be done:

- a mathematical description has to be formulated,
- it has to be analyzed at the starting point,
- the desired form of the objective function ψ has to be specified,
- the optimization variables have to be selected,
- the desired form of the constraining functions q_j has to be specified,
- a suitable optimization method has to be selected,
- the resulting mathematically formulated optimization problem has to be solved, and
- using the mathematical model, the results have to be transformed back into the dynamic model (for dynamic problems only).

General Aspects of the Optimization Process

If the aim of an optimization process is to optimize several properties that simultaneously affect the system (such as minimizing the size values while respecting the electrical properties), we obtain a multi-criteria objective function. The objective function then takes the form of a weighted sum of single-criterion

functions. Each of these functions generally assumes its local minima at different points of the optimization parameter spaces. This is the reason why a multi-criteria function can have a large number of shallow local minima or is insensitive to changes in the optimization parameters. Due to this fact, the selection of an optimization method is of great importance. The result is averaged in the sense that several criteria may participate simultaneously in a reduction of the multi-criteria function, while some other criteria may increase.

A more suitable method may be to select a single-criterion objective function, including all criteria in the constraints. Only the most significant criterion is chosen for the objective function to be specified in the subsequent process. All other criteria included in the constraints are kept within specified limits without being optimized. Thus, the results of an optimization process are dependent on the degree of reduction of the admissible set given by the inequality-type constraints.

Generally, we specify the constraints in a form similar to the objective function

$$q_i(p) = f_i(p) - f_i^h, \quad i = 1, 2, \ldots, m^* \tag{34.1}$$

Here f_i are suitable functions of a vector variable and f_i^h their maximum admissible values.

The selection of optimization variables is given by the sensitivity of the objective function to changes of relevant optimization variables. This sensitivity is described by the gradient vector of the objective function.

$$\text{grad } \psi(p) = \left[\frac{\delta\psi(p)}{\delta p_1}, \ldots, \frac{\delta\psi(p)}{\delta p_s}\right]^T \tag{34.2}$$

Types of Optimization Methods

Standard Optimization Methods

Most practical problems lead to nonlinear (transcendental) systems of equations. These may only be solved using numerical optimization methods. According to the order of the derivatives used in the application of a method, numerical methods of finding local minima of functions of several variables may be divided into:

1. zero-order methods (comparative)
   - methods of co-ordinate comparison
   - simplex methods
   - stochastic methods
2. first-order methods (gradient and quasi-gradient)
   - methods of associated directions
   - variable-metric methods
3. second-order method (Newton method)

Stochastic Methods

These methods consist of calculating the values of the objective function at a large number of selected points. The points are selected by such criteria that each point in the space has an equal probability of being selected. The best points are then determined by comparing the function values. From the outlined strategy, it follows that these methods lead to computing the function values at a large number of points, which may protract the calculation. On the other hand, we can more easily reach the global optimum of the function to be optimized. These methods also comprise the evolution methods since the first solution population is generated completely by random. The difference only consists in the strategy of selecting better solutions.

Evolutional Optimization Methods

Since some problems are difficult to solve by standard numeric optimization methods, even if they converge to an acceptable optimum in a reasonable time, new approaches had to be developed. Therefore, a number of new methods have been designed based on the laws of natural genetics copying them in various degrees. These methods employ random generation of input parameters and so can be thought of as stochastic methods. In general, stochastic optimizing algorithms (including virtually all the evolutional algorithms) optimize using multi-parameter function with "wild" behavior, that is, with many minima or with an unknown gradient. Stochastic optimization methods are necessarily slower than heuristic approaches, which take advantage of the fact that they know the type and details of the function to be optimized. Unless the conditions for the global optimum are previously known, we can never be sure whether we have reached the global optimum to be able to terminate the optimization process. However, stochastic optimization methods also bring numerous benefits. They are generally very well specified and thus applicable virtually to any problem, and they can get out of the trap of a local minimum. The evolutional process of searching the space of potential solutions requires an equilibrium of two objectives:

- to find the nearest (mostly local) minimum as quickly as possible, and
- to search the space of all potential solutions in the optimum manner.

The methods differ in their orientation towards these two objectives and they can be roughly ordered in a sequence starting with methods tending to local minima to methods searching a large number of potential solutions:

1. Stochastic "hill climbing" algorithms,
2. Tabu search algorithms,
3. Simulated annealing algorithms, and
4. Genetic algorithms.

Hill Climbing Algorithm

This is the simplest optimization algorithm being a variant of the gradient method "without gradient" where the direction of the steepest climb is determined by searching the neighborhood. This algorithm also has all the drawbacks of gradient methods, in that it is very likely to end up in a local extreme without reaching the global minimum. Here the starting solution is generated at random. For the currently designed solution, a certain neighborhood is generated using a finite set of transformations and the best minimum is chosen from this neighborhood. The local solution obtained in this way is then used as the center of a new neighborhood in which the optimization is repeated. This process is iterated a specified number of times. In the course of this process the subsequent best solutions are recorded to be finally used as the resulting minimum. The basic drawback of this algorithm is that, after a number of iterations, it may revert to a local minimum that has already been passed in a previous step (the problem of looping). This problem can be avoided by running the algorithm several times with different randomly generated initial values to eventually choose the best result achieved.

Tabu Search Algorithm

At the end of the 1980s, Professor Fred Glover designed a new approach to solving the problem of finding the global minimum, which he called *tabu search*. At present, this method is among those used most frequently to solve combinatorial problems and problems of finding the global minimum. Based on the hill-climbing algorithm, it tries to eliminate the problem of looping. The hill-climbing algorithm is equipped with a so-called short-time memory, which, for a short previous interval of the algorithm history, remembers the inverse transformations to locally optimal solution transformations used to obtain the new centers in iterations. These inverse transformations are prohibited (tabu) when the new neighborhood is created for a given current solution. In this way, the looping caused by falling into the trap of a local minimum may substantially be reduced. A hill-climbing algorithm modified in this way systematically searches the entire area in which the global minimum of a function is to be found.

Simulated Annealing Algorithm

Apart from the stochastic methods and methods based on natural evolution, there is another possibility of simulating the evolution of systems based on the physical evolution of macroscopic systems. The annealing of a solid body in order to remove the internal stress is a simple example of this kind of evolution. For a physical interpretation of this process, consider a body that is heated until it reaches a high temperature. The temperature is then gradually lowered. The atoms of a body heated to a high temperature can easily overcome the local energetic barriers to reach equilibrium states. When the temperature is lowered, atoms are fixed in this state and the cooled off body is without internal stress.

This principle was used to design the method of simulated annealing. First, an initial temperature T_{max} is set, whose value is important for the method to be efficient. The simulated annealing algorithm then searches the space of all potential solutions in a strongly stochastic way, also accepting the states that correspond to solutions worse than the current one. This property of simulated annealing is a characteristic feature of this method and provides a way of escaping from a local minimum trap, thus allowing the search of another area of the entire solution space. However, as the annealing temperature T is lowered, the probability of accepting worse states as well is diminishing. For small temperature values then, only solutions better than the current one are considered.

Genetic Algorithm (GA)

Genetic algorithms (GAs) are most frequently used to optimize the parameters of an unknown system whose mathematical description is either too complicated or unknown [5]. When applying a GA, it is mostly sufficient to know a function assigning a price to each individual in the population. This may be the error of the solution for randomly selected parameters during GA. Since a GA is looking for a maximum, the error, which, on the contrary, is being minimized, must be transformed into looking for a maximum. This may be done in several different ways: by subtracting the error from the maximum error occurring, by calculating the inverted value of the error, or by using another transforming function that approaches zero as the error approaches one. Increased attention should be paid to setting up the program implementing the pricing function since it consumes the most computing time compared with the other GA components.

Apart from general optimization problems, GAs are mostly applied to neural networks. Here the tendency is to employ GAs at two different levels. First, for finding suitable weights for a neural network and second, when optimizing the structure of a neural network, that is, when selecting the algorithm, the number of input neurons in the hidden layers, the number of hidden layers, etc. Using a genetic algorithm to optimize the parameters of another genetic algorithm (the size of the population, the number of crossbreedings, the extent of mutations, the frequency of mutations) is a very revolutionary idea (optimization of the computation time where the computation time is a pricing function of the GA). As far as applications of GAs to problems encountered in research of electric machines are concerned, GAs have been used to identify the parameters of the substitution diagram of an induction motor.

By way of conclusion, it may be added that genetic algorithms perform surprisingly well when all other algorithms fail, such as for incomplete problems where the computation time is an exponential or factorial function of the number of variables. There is no point in using GAs to optimize relatively simple functions or functions for which special algorithms exist for their description. Considering the necessity to calculate the function values for tens or hundreds of genetic chains in a population and the necessity to evaluate hundreds or even thousands of populations during a single run of the program, GAs are rather time-consuming.

Despite the positive results achieved by using GAs, it is clear that nature must use even more intricate and, at the same time, not very sophisticated methods. The GAs described above only correspond to very primitive examples observed in nature, particularly those related to asexual reproduction with a single chromosome. Since nature has taken billions of years to test its algorithms, it is highly efficient to further learn from it. It is interesting that it needs no mathematics to solve complicated problems of optimization. Nevertheless there are other optimization methods suitable to solve the problems of the design [2–4].

Selection of a Suitable Optimization Method

The standard gradient method is still one of the most frequently used methods. Gradient methods or even the standard non-gradient methods (such as the simplex method) are not suitable if the finding of the global minimum is required of a function with many local minima. Mostly, these methods only reach an insignificant minimum close to the starting point (the initial solution) in which they are trapped. This deficiency is mostly removed by repeatedly selecting at random the initial solution of an optimization problem and taking the best result for the solution. The stochastic character of this process can only be seen in the random selection of the initial solution. The subsequent optimization algorithm then proceeds without any randomness. Then the evolutional optimization methods are thought of as stochastic ones despite their employing of a certain strategy when choosing the getter points. The following are the main differences between a genetic algorithm and the more frequently used gradient method:

- GA performs no gradient computation, which might be difficult and time consuming particularly for large systems, and
- GA works with randomly generated solutions and may converge more quickly to the global minimum.

To optimize the draft design of an induction motor, an optimization method was employed using a genetic algorithm. This method is described in more detail in the following chapter.

34.3 Optimum Design of Induction Motor (IM)

IM Design Introduction

Actual design of an induction motor usually depends on the requirements of individual customers, who specifically define parameters which a designed machine should accomplish. In this way, with the same machine output, we can obtain different implementations that meet individual conditions more or less. It is possible to require a good quality of one parameter only with the deterioration of other parameters. We are going to deal with a design of motors of (0,6–200) kW outputs. Motors are designed for permanent load and with the project assignment the following input values are required:

Machine output P_n [kW], voltage U_{1n} [V], winding connection *Y*/*D*, number of poles 2**p** or rotation speed n [min^{-1}], grid frequency f [Hz], efficiency η[%], power factor cos φ, insulation class, IP implementation, and the shape of the machine.

We consider squirrel-cage motors in closed implementation with framework and cooling ribs. There is a cast aluminum rotor cage. For the design, data such as conductors and slots dimension or magnetic characteristics deducted from tables and graphs that are given by the standard or by the manufacturer's measurement, are needed. The actual design is a compromise between individual design parameters, so that a resulting machine would have the best possible operating characteristics with a perfect heat and material utilization. The actual motor design is described in the following section.

Classical IM Design Evaluation

An induction motor design, when carried out manually, represents hundreds of calculations, which can last tens of hours even with an experienced constructor. As computers made their way into practically all branches of design and analysis, a series of programs which co-operate with a designer in an interactive fashion and speed up a calculation were created.

In spite of indisputable advantages of this design process, we have to realize that there is a remarkable quantity of various design implementations of the given motor which, more or less, achieve the required motor operating characteristics. This approximates the global minimum of an objective function, which evaluates the design quality.

Thus, the idea to create a program for searching the whole state space of all possible solutions and selecting such a variant, which is the most appropriate to an evaluating objective function (the required

TABLE 34.1 Generated Parameters List and Setup of Their Limits

Parameter Name	Symbol	Dimension	High Limit	Low Limit
Stator outside diameter	D_e	mm	User optional	User optional
Stator inside diameter	D	mm	User optional	User optional
Ideal iron length	l_i	mm	User optional	User optional
Air gap induction	$B\delta$	T	0.5	1.0
Stator slot filling	k_{dr1}	—	0.6	0.75 (0.8)
Air gap size	δ	mm	0.2	0.4
Stator current density	σ_1	A mm^{-2}	3.0	15.0
Rotor rod current density	σ_t	A mm^{-2}	2.0	6.0
Rotor ring current density	σ_k	A mm^{-2}	2.0	4.0
Teeth magnetic induction	B_z	T	1.6	2.0
Slot number per pole and stator phase	q_1	—	2.0	5.0

motor characteristics) using some of the optimization methods, was developed. The stochastic evolutionary method genetic algorithm was selected, because it searches the whole state space of all possible solution in a best way.

Description of a Solved Problem

Generated Parameters

The values given in Table 34.1 are recommended only, and, for the motors of outputs below 200 kW, they are mostly limiting values. Varying the parameter values or substituting one parameter for another is possible only by intervention in the program source text. Diameter limits D_e and D from the input file are considered only in the case that a motor design without regard to standardized axis height is required. In the case that standardized axis height is entered, these limits are calculated. Limits of an ideal rotor length are appropriate to enter as narrow as possible for faster convergence to a limit. But this is not a required condition. Generally, the lower the range of individual parameters, the faster the convergence to a global minimum, and the number of local minimums is lower.

Objective (Criterion) Function

It is not just the form of the objective function, but also the selection of optimized parameters that is important for good optimization results. Selected parameters must sufficiently describe a quality solution to the given problem. In the case of an induction motor design optimization task, the following parameters were selected:

Motor volume	V [dm^3]
Motor temperature rise	ϑ_n [K]
Motor nominal power factor	$\cos\varphi_n$ [−]
Motor nominal efficiency	η_n [−]
Torque overload capacity	m_{pn} [−]

These parameters are most important for the quality of the design and describe the design sufficiently.

The total error is based on the relation given in Eq. (34.1), a sum of individual partial errors of each controlled parameter. If we put more emphasis on some parameter, we increase a corresponding weight coefficient, thus achieving its improvement in the final design. At the same time, values of other parameters will decrease. Finding an optimal setup of gain coefficients is one of the most important and difficult problems. The term "optimal setup" means that the designed motor has the highest power factor, efficiency, and torque overload capacity values at the minimal volume and simultaneously does not exceed a permitted temperature rise for a selected insulation class. We have the relationship

$$\begin{aligned}\varepsilon(\mathrm{GR}_i) = {} & \mathrm{abs}(kV \cdot V) + \mathrm{abs}(k\vartheta(0.89\,\vartheta_d - \vartheta_n)) + \mathrm{abs}(k_{\cos\varphi}(1-\cos\varphi_n)) \\ & + \mathrm{abs}(k\eta(1-\eta_n)) + \mathrm{abs}(k_{mp}(m_p + 1 - m_{pn}))\end{aligned} \qquad (34.1)$$

TABLE 34.2 Input Values of 5.5 kW, 380 V Motor

Quantity Name	Symbol	Dimension	Value
Nominal power output	P_n	W	5500
Nominal voltage	U_{1n}	V	380
Required power factor	cosj	—	0.81
Required efficiency	η	—	0.86
Grid frequency	f	Hz	50
Motor axis height	H	mm	132
Number of pole pairs	p	—	3
Temperature class of insulation	TT	—	F
Torque overload capacity	m_p	—	2

where kV is the volume weight coefficient, $k\vartheta$ is the temperature factor weight coefficient, $k_{\cos\varphi}$ is the power factor weight coefficient, kh is the efficiency weight coefficient, k_{mp} is the torque overload capacity weight coefficient.

Achieved Results

5.5 kW, 380 V Motor Design Description

During program development and tuning, an optimization was performed on the motor described in Table 34.2. In this section we describe the results and problems encountered in the optimization process. The symbols and quantities, which are not explained in detail, were either used in the previous text or are listed at the list of used quantities at the beginning of the document. The motor input parameters are given in Table 34.2.

Other Results

It follows from the physical principle that optimized quantities are closely related. Increase of a gain of one quantity results in a disadvantage of the quantities. Based on the performed optimizations, it can be concluded that two kinds of motors exist, depending on the content of iron and copper:

1. Motor with prevailing iron content, high stator current density, good power factor, for a price of worse efficiency of the motor and with slightly worse torque overload capacity than the second motor.
2. Motor with high copper content and conversely low stator current density, good efficiency, and worse power factor value. Torque overload capacity is good.

The type of motor is determined based on the setting of gain coefficients. A sum of temperature and power factor errors on one side impacts the sum of volume and efficiency errors on the other side. The torque overload capacity can be good for both kinds of motors.

Results of individual optimizations are listed in Table 34.3, ordered by volume value from smallest to largest. Different varieties of motors were obtained, depending on values of gain coefficients. It is difficult to determine which solutions are good or bad, because the selections depend on actual customers' requirements. The solution that gives the best value of optimized quantity is marked in bold. Solution numbers 1, 5, 8, 23, and 25 can be considered successful from this perspective. The motor described above (solution no. 2) serves for depicting of the task. Previously-mentioned relations between individual quantities can be observed in Table 34.3, which lists the optimization results without limiting the generated parameter, thus using the requirements in Table 34.1.

Next, a motor optimization was performed with just one optimized parameter, when the gain of the other parameters was set equal to zero.

1. *Volume optimization*. In this case, the algorithm selected as the best solution motors with minimal dimensions, when parameters D_e, D, and l_i were converging to minimum preset limits.
2. *Temperature optimization*. The algorithm reached first reached a local minimum with temperature at the maximum based on the required insulation class, and mostly stayed on this value.

TABLE 34.3 Motor $P = 5.5$ kW, $U = 380$ V Solutions List, Without Generated Parameters Limited

Number	V [dm^3]	ϑ [K]	$\cos\varphi$ [-]	η [-]	m_p [-]	Directory
1	**3.96**	88.1	0.798	0.834	1.72	Motor1
2	4.20	86.9	0.818	0.843	1.90	Motor2
3	4.31	74.9	0.787	0.865	1.77	Motor3
4	4.32	88.8	0.836	0.817	1.78	Motor4
5	4.33	75.1	0.690	**0.973**	1.07	Motor5
6	4.50	89.0	0.836	0.834	1.79	Motor6
7	4.51	86.8	0.818	0.818	1.93	Motor7
8	4.54	**90.0**	0.884	0.812	1.98	Motor8
9	4.56	84.6	0.857	0.816	1.74	Motor9
10	4.58	86.5	0.836	0.817	1.77	Motor10
11	4.63	68.2	0.792	0.858	2.10	Motor11
12	4.69	88.4	0.862	0.808	1.80	Motor12
13	4.70	73.4	0.845	0.830	2.25	Motor13
14	4.73	61.0	0.799	0.871	1.90	Motor14
15	4.78	78.1	0.853	0.858	1.67	Motor15
16	4.78	71.0	0.767	0.870	1.80	Motor16
17	4.81	70.6	0.703	0.934	1.28	Motor17
18	4.97	54.5	0.804	0.883	1.90	Motor18
19	5.08	55.5	0.762	0.877	2.20	Motor19
20	5.12	88.2	0.879	0.806	2.05	Motor20
21	5.96	44.0	0.784	0.870	2.55	Motor21
22	6.35	42.4	0.803	0.882	2.69	Motor22
23	6.40	87.5	**0.887**	0.853	2.27	Motor23
24	6.57	59.0	0.747	0.956	1.16	Motor24
25	7.05	42.3	0.793	0.865	**3.00**	Motor25
26	7.39	52.9	0.714	0.986	1.03	Motor26

3. *Power factor optimization*. An effort to achieve the first type of motor (see above discussion) with low copper content, high current density σ_1, worse value of efficiency. Torque overload capacity was good.
4. *Efficiency optimization*. The designed motor corresponded to the second type of motor (see above discussion) with prevailing copper content, low current density σ_1, and good efficiency, however, with worse power factor values. Torque overload capacity was good.
5. *Torque overload capacity optimization*. The motor is designed with high number of slots for pole and phase, resulting in gradual spread of conductors on the perimeter. The motor can have prevailing iron or copper content depending on a local solution, to which it converged. It can have good values of power factor and efficiency for a price of machine volume increase.

34.4 The Use of a Neuron Network for the Identification of the Parameters of a Mechanical Dynamic System

The basic step used to solve the dynamic tasks by means of any type of modeling is to create a set of important quantities that include both the quantities describing structure, conditions, and the interactions of technical objects and the quantities that characterize the consequences (i.e., their demonstration and behavior).

The methods of creating the mathematical models in drive systems, in general an interactive process, utilize

- the applications of well-known physical principles that describe the phenomena in drive systems (e.g., the second Newton's principle, Kirchhoff's laws, etc.), or
- the applications of the methods based on artificial intelligence algorithms (e.g., genetic algorithms [1] and artificial neuron networks [6, 7]).

The theories on which the methods of artificial intelligence are based replace the “standard” analytical and numerical methods when

- these are the only theories which can solve the problem,
- they exhibit better properties from the point of view of the problem solution (e.g., a better conditionality considering the changes of input values), and
- they allow the problems to be solved more effectively.

The last case is typical when we want to approach the real operational conditions as close as possible. From various methods of artificial intelligence, the stochastic evolution algorithms and the artificial neuron networks are being increasingly utilized in the field of the modelling of drive interactive systems. In the following section, two methods are shown that are applied to the problems of the analysis of dynamic properties in drive systems.

The solution of dynamics by means of the algorithms of artificial intelligence represents a solution of the following partial problems:

- Specifying the set of important (relevant) quantities,
- Selecting the theory which is suitable to solve the problems,
- Arranging the relations among relevant quantities so that they allow the selected algorithm of artificial intelligence to be used,
- Generating the training data, and the selection of the method of teaching, for example, in neuron networks, and
- Testing the quality of the results reached and their evaluation.

Practical Application

Many identification methods are known and very often verified quite well in practical terms. The limit factors that make these procedures more difficult (e.g., the assumptions about system linearity, stationarity, and normality of the phenomena which occur in the systems, etc.) are also known. Hence, we have used an untraditional approach to the problem *f* identifying the dynamic properties in mechanical systems for which the use of neuron systems seems to be promising, and at the same time available for engineers' thinking.

A Practical Application—Gearbox

The first case that was analyzed is a vehicle gearbox. Inputs (engine load), outputs (frequency-amplitude spectra of torsional oscillations), and the gearbox structure were known. The ilete information was related to the selected parameters of stiffness and damping in the drive. Due to variable operational conditions, the magnitude of damping may vary enough to significantly affect output characteristics. If many experimental results are evaluated, some typical failures can be identified and their possible occurrence can be anticipated from output files. We have used the data that were measured in the real system. The frequency-amplitude spectrum of torsional oscillations was measured at the gearbox shaft, which is seated on four bearings with five speed gears (Fig. 34.1). Originally, the measurement was carried out to determine the resonance of frequency systems with the goal to reduce noise. The following parameters were set in the system:

Stiffness $K \in \{0.3, 12.0, \bullet\}$,
Damping $B \in \{-, 0, 0.3, 12.0\}$.

The measurements were done with testing frequencies $f = 512$ Hz and $f = 1024$ Hz. To record significant oscillation harmonics, the excitation frequency was flexible in both cases:

1. from 2.5 to 14.0 Hz in steps of 0.5 Hz, and
2. from 14.0 to 40.0 Hz in steps of 1.0 Hz.

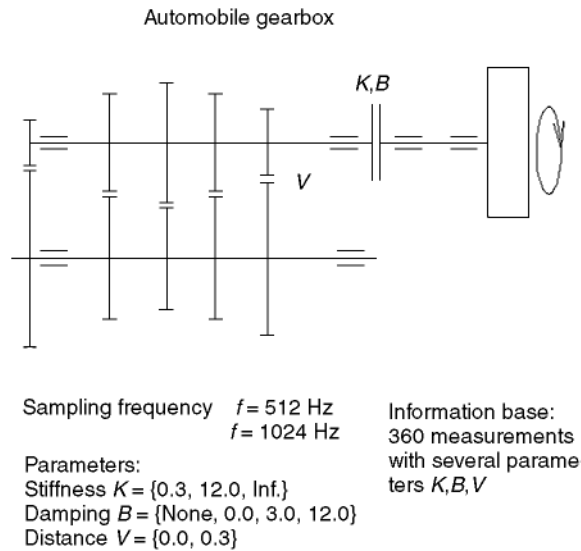



FIGURE 34.1 Five speed gear.

TABLE 34.4 Expected Natural and Excitation Frequencies for the Gearbox

Table of Expected Frequencies		[Hz]
Low frequencies		up to 5.0
Operational frequency (OF) (speed)	OF	14.16
	2 ¥ OF	28.32
	3 ¥ OF	42.48
Interharmonic frequency (IHF)	0.5 ¥ OF	7.08
	1.5 ¥ OF	21.25
	2.5 ¥ OF	35.42
Natural frequency (NF)	I.NF	43.91
	0.5 ¥ I.NF	21.96
	2 ¥ I.NF	87.82
	II.NF	322.1
	0.5 ¥ II.NF	161.1
	2.0 ¥ II.NF	644.2
Combination frequency (CF)	2 ¥ OF + 0.5 ¥ I.NF	50.28
	OF + I.NF	58.07
	2 ¥ OF + I.NF	72.23
	2 ¥ OF + 2 ¥ I.NF	116.1
Tooth frequency (TF) {TF = z.OF}	1st speed gear	184.1
	2nd speed gear	325.7
	3rd speed gear	424.9
	4th speed gear	580.6

The spectrum always included 512 spectral lines. The measurement was repeated 360 times for different variations of the parameters K and B (the system adjustment). The values of natural and excitation frequencies, which are expected for the gearbox to be analyzed, are shown in Table 34.4.

Task Definition

The task was originally defined in the following way: to estimate the corresponding magnitudes of parameters K and B (this means, to recognize the adjustment of parameters used in the mechanical dynamic system) by means of the artificial neuron network on the basis of the frequency-amplitude

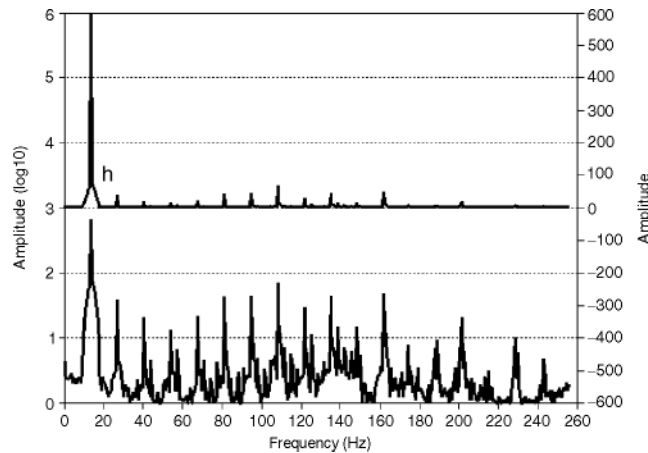



FIGURE 34.2 Stimulus vectors normalization (top, before normalization; bottom, after normalization).

spectrum of torsional shaft oscillations in the given system. A multilayer perceptron with three layers (i.e., input, hidden, and output) is used as the configuration of the neuron network. Select the signal neuron functions as the linear, hidden, and output layers of the logistic function, $f(x) = 1/(1 + e^{-x})$.

According to this definition, this is to identify the system parameters on the basis of the measured frequency-amplitude spectra. However, the parameters are taken from discrete sets (and very low), and the task could be redefined as the "standard" task of the spectrum classification according to seven attributes (each attribute corresponded to one of the possible values of the parameters *K* and *B*). The application of neuron networks to solve such a problem is more successful when compared to the solution of the original task.

The amplitudes of spectral lines were expressed in logarithm scale, and a reduction of spectral dynamics with an increase of their informative quality has been achieved. Considering the nonlinear nature of the activation neuron functions used, which extends beyond the saturation range for the input interval ⟨0.5, 0.95⟩, the network cannot respond well to stimulous vectors with a high range of the values in the individual components. This is illustrated in Fig. 34.2. The input network layer was configured to 512 input neurons. The amplitude logarithmic value of one spectral line was entered into each input. The individual neurons in the output layer correspond to the classification attributes. Because there are seven attributes, seven neurons were configured in the output layer.

The only-hidden layer was set as the arithmetic mean of the number of input and output neurons. Two hundred sixty neurons were configured to the only-hidden layer, as illustrated in Fig. 34.3. Each item corresponded to one measurement of the frequency spectrum (a stimulus vector) with a corresponding attribute vector (a vector of the required responses). The specific variation of the parameters was expressed by the required network response to two corresponding output neurons equal to 1, and the remaining output neurons equal to 0.

From the original 360 items, 36 items were randomly separated (10% of total) for the future tests. We ensured that the network tests would be carried out with the items that have not passed the training network process (the network was not trained to these situations). This is necessary to verify the generalization model properties. The training set was formed by the remaining 324 items. The sequential strategy of teaching was used, i.e., the items from the training set were used in the teaching process with the fixed sequence (cyclic passages through the training set). Taking into consideration the size of the neuron network to be configured, the method of feedback moment propagation has been selected as the teaching method that exploits only the information up to the first-order inclusively (the values of a special function—a teacher and his gradient), and it has not used Hessian or its estimate, which, in this case, would be very demanding.

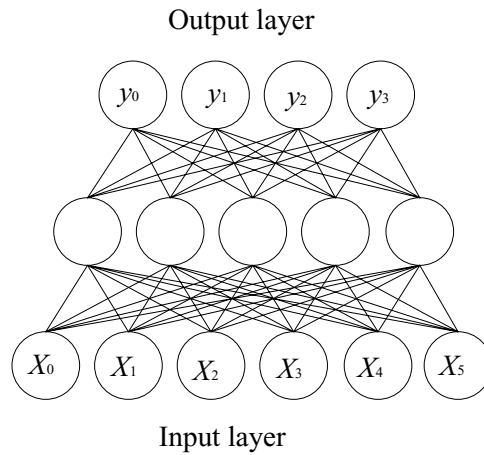



FIGURE 34.3 Network configuration (hidden layer contains neurons without labels).

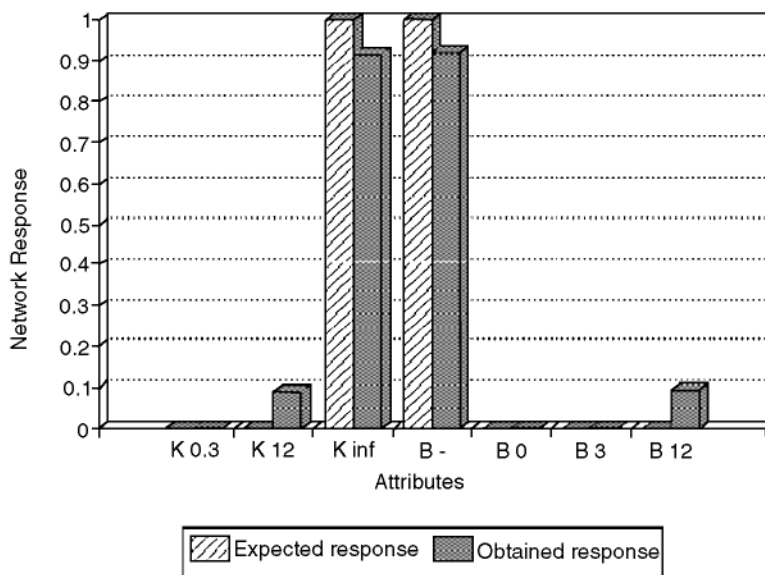



FIGURE 34.4 Successful response of neural net.

Results

The neuron model of the mechanical system has manifested a high rate of success during the verification by test sets. The network was taught with random selections of the test items. During testing of the individual models, the responses of the network were successful in 85–95% of all cases (see Figs. 34.4 and 34.5). Moreover, the estimate of the values K and B that correspond to the parameters is available within a couple of seconds for the frequency spectrum in the active mode. However, it is possible that a model with higher quality will be achieved if special optimizing techniques are used in the future.

In summary, the neuron model of the mechanical system described above can be assessed as usable in practical terms.

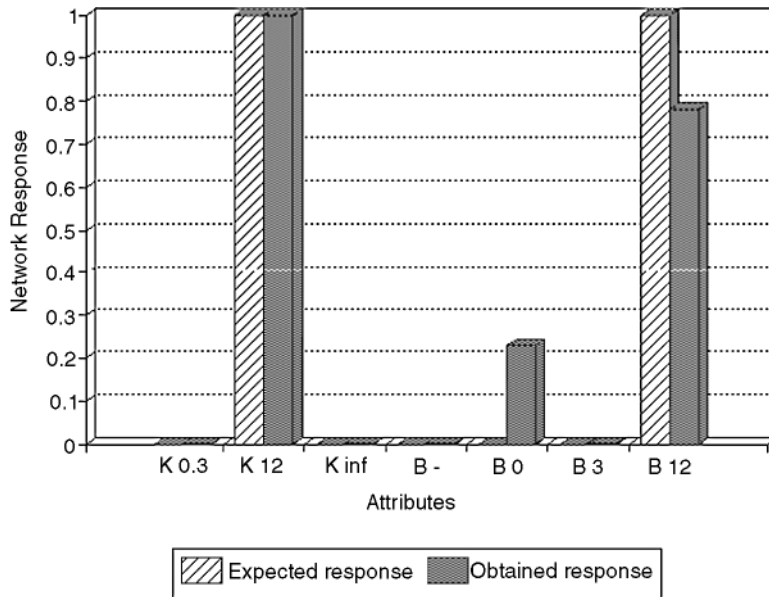



FIGURE 34.5 Failure response of neural net.

References

1.Goldberg, D., *Genetic Algorithms in Searching, Optimisation and Machine Learning,* Reading, MA: Addison-Wesley, 1989.

2.Glover, F., Lagunai, M., Marti, R., Fundamentals of scatter search and path relinking, *Control and Cybernetics,* pp. 653–684, 2000.

3.Glover, F., *Scatter Search and Star-Paths—Beyond the Genetic Metaphor,* pp. 125–137, New York: Springer-Verlag, September 1995.

4.Glover, F., Kelly, J.P., Langunai, M., Genetic algorithm and tabu search—hybrids for optimization, *Computers and Operations Research,* pp. 111–134, January 1995.

5.Lee, J., Hajela, P., Parallel genetic algorithm implementation in multidisciplinary rotor blade design, *Journal of Aircraft,* Vol. 33, No.5, pp. 962–969, September–October 1996.

6.Hagan, M.T., Demuth, H., Beale, M., *Neural Network Design,* Boston: PWS Publishing, 1996.

7.Kosko, B., *Neural Networks and Fuzzy Systems,* Englewood Cliffs, NJ: Prentice-Hall, 1992.

8.Ye, X., Loh, N., Dynamic system identification using recurrent radial basis function network, *Neural Networks Theory, Technology, and Applications,* New York: IEEE Technology Update Series, 1996.